THREE ESSAY IN ECONOMETRICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sida Peng

May 2017

THREE ESSAY IN ECONOMETRICS

Sida Peng, Ph.D.

Cornell University 2017

The development of statistical tools benefit economic modeling by relaxing the assumptions for identification. For example, kernel or SEIVE methods allow us to remove parametric assumptions on the functional form. Penalized estimators allow us to run a regression when the dimension of the data exceeds the number of observations. This dissertation presents new economic models and new estimators based on LASSO-type estimators and kernel estimators. I study their statistical properties and show how these new estimators can be applied to study specific economic problems.

The first chapter, *Heterogeneous Endogenous Effects in Networks*, proposes a new method to identify leaders and followers in a network. Current literature models peer effects using spatial autoregression models (SARs). SARs implicitly assume that each individual in the network has the same endogenous effects on others and conclude the key player in the network to be the one with the highest centrality. However, when some individuals are more influential than others, centrality may fail to be a good measure. I develop a SAR model that allows for individual-specific endogenous effects and propose a two-stage LASSO (2SLSS) procedure to identify influential individuals in a network. My method allows me to identify leaders and followers through their observed behaviors on the network. Under an assumption of sparsity: only a subset of individuals (which can increase with sample size n) is influential, I show that my 2SLSS estimator

is consistent and achieves asymptotic normality. I develop robust inference including uniformly valid confidence intervals. These results also carry through to scenarios where the influential individuals are not sparse. I extend the analysis to allow for multiple types of connections (multiple networks), and I show how to use the square-root sparse group LASSO to detect which of the multiple connection types is more influential. Simulation evidence shows that my estimator has good finite sample performance. I further applied my method to the data in [9] and my proposed procedure is able to identify leaders and effective networks.

The second chapter, *On Testing Continuity and the Detection of Failures*, propose a new estimator to detect discontinuities in the functional form. This is coauthored work with Professor Matthew Backus. Estimation of discontinuities is pervasive in applied economics: from the study of sheepskin effects to prospect theory and the bunching" of reported income on tax returns. This salience of discontinuities makes the models that generate them empirically testable. However, detection and identification of those discontinuities typically relies on knowledge of their number, their type, their location, or their underlying functional form. We develop a nonparametric approach to the study of arbitrary discontinuities –point discontinuities as well as jump discontinuities in the nth derivative, where n = 0,1 that does not require ex ante knowledge of their number or location. Our approach exploits the recent development of false discovery rate control methods for LASSO regression as proposed by [46]. This framework affords us the ability to construct valid tests for both the null of continuity as well as the significance of any particular discontinuity without the computation of nonstandard distributions. We illustrate the method with

a series of Monte Carlo examples and by replicating prior work, e.g. classical regression discontinuity election study [60], [24] and [6].

The third chapter, *Local Regression Smoothers with Set Valued Outcome Data*, provides statistical results on local linear regression smoothing when the outcome data is set valued and the regressors are exactly measured. This is coauthored work with Qiyu Li, Professor Ilya Molchanov and Professor Francesca Molinari. We derive the asymptotic properties of our estimator, propose a bias correction method, and adapt results from [15] to obtain point-wise confidence bands that asymptotically cover the functional of interest with probability 1-α. We demonstrate the usefulness of our approach using a novel dataset that follows 132 patients during anti-cancer treatment.

BIOGRAPHICAL SKETCH

Sida Peng received his B.A. in economics and mathematics (with distinction) and M.S. in statistics in May 2011 from University of Virginia. In August 2012, he began his graduate studies in Department of Economics at Cornell University. During his study, he was supported by Sage Fellowship (2012), Jacobs Fellowship (2015), Ruth Ada Birk Eastwood Fellowship (2015), Mabel A. Rollins Scholarship (2015) and Ernest Liu 64, Ta-Chung and Ya-Chao Liu Memorial Fellowship (2016). He received his M.A. in economics in January 2016 and will receive his Doctor of Philosophy Degree in May 2017.

To my beloved parents Xin Peng and Jingxiang Shi.

and

To my beloved wife Yanlei Ma

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest thanks to my committee chair, Professor Francesca Molinari, for her continuous support of my Ph.D study, for her guidance, patience, and encouragement. I could not have imagined having a better advisor and mentor for my Ph.D study. I have learned a lot from working with Professor Molinari, not only in terms of intellectual knowledge, but also how to live, work and think as a researcher.

I would like to thank the rest of my thesis committee: Professor Matthew Backus, Professor David Easley, and Professor Marten Wegkamp, for their insightful comments and questions. Their suggestions and criticisms are invaluable in helping me complete the three chapters in this thesis.

I would also like to thank Professor Donald Kenkel and Professor Zhuan Pei, for their great support on my research and on my job market. I could not achieve what I have at this moment without them.

I would like to thank my fellow doctoral students for their feedback, proofread and of course friendship.

My acknowledgement also goes to the Ta-Chung and Ya-Chao Liu Memorial Fellowship, Mabel A. Rollins Scholarship, Ruth Ada Birk Eastwood Fellowship, Jacobs Fellowship, and Sage Fellowship. These generous financial supports help me to concentrate on my research and make my life much easier.

Finally, I would like to express my deep and sincere gratitude to my family.

I am grateful to my wife Yanlei Ma for her love, for her tolerance of my occasional bad mood, for her patience when preparing me for the job market, for her encouragement and support. I am deeply indebted to my parents Xin Peng and Jingxiang Shi. I couldn't accomplish anything without their love, support and help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

HETEROGENEOUS ENDOGENOUS EFFECTS IN NETWORKS

1.1 Introduction

How an individual's behavior is affected by the behavior of her neighbors in an exogenously given network is an important research question in applied economics. With the increasing availability of detailed data documenting connections among individuals, spatial autoregression models (SARs) have been widely applied in empirical networks literature to estimate endogenous effects.

In SARs, an individual's behavior depends on the weighted average of other individuals' behaviors [see 5, 57]. Standard SARs assume that the endogenous effects are the same across individuals in a network. Each individual influences her neighbors *at the same rate* regardless of who she is. However, in many contexts, some individuals are clearly more influential than others. For example, [76] finds that the magnitude of spillovers varies dramatically among workers with different skill levels. [27] also note that popular teenagers in a school have much stronger influence on their classmates' smoking decisions than their less popular peers.

I propose a novel SAR model which allows for *heterogeneous* endogenous effects. Each individual in a network simultaneously generates an outcome that takes into account all her neighbors' behaviors. Unlike standard SARs, each individual has an individual-specific effect on her neighbors. As a result, there are as many coefficients for individual-specific endogenous effects as there are individuals in the network. To achieve identification, I assume that "truly-

influential" individuals only constitute a small fraction of the total population. In other words, individual-specific coefficients are assumed to be sparse. This assumption allows me to estimate the model via the least absolute shrinkage and selection operator (LASSO). The LASSO procedure penalizes the l_1 norm for the coefficients of heterogeneous endogenous effects. The geometry of the l_1 norm enforces the sparsity in the LASSO estimators. If a coefficient is selected by LASSO (i.e. the estimated coefficient is non-zero), the individual associated with this coefficient can influence all her neighbors at her specific rate. Otherwise the LASSO estimator will indicate that the individual has no influence on her neighbors. With some restrictions on the network structures, I show that the LASSO estimates for heterogeneous endogenous effects have near oracle performance [see 21]. In other words, the selection of influential individuals is consistent and the convergence rate of non-zero LASSO estimates is the same as the convergence rate that would have been achieved if the truly influential individuals were known.

One challenge in my estimation process is the presence of endogeneity in the spatial lag and error term. As with standard SARs, the dependent variable in my model is used to construct spatial lags as an independent variable. As a result, the regressors are correlated with the error term and estimates would be biased if we were to apply LASSO directly.

First I propose a set of novel instruments to address the endogeneity. Following [57], I express the dependent variable as an infinite sum of functions consisting of independent variables and an adjacency matrix. These functions are used as instruments. Then I design a two-stage estimation process for heterogeneous endogenous effects using LASSO at each stage. *In the first stage*,

I use LASSO to estimate the coefficients for the instruments. These estimated coefficients and instruments are then used to create a synthetic dependent variable. *In the second stage*, I replace the dependent variable in the spatial lags with the synthetic variable to perform the LASSO estimation. Unlike in the standard two stage least square estimation process, the synthetic dependent variable in the first stage suffers from a shrinkage bias due to the LASSO fitting. However, I show that with certain restrictions on the network structure, the shrinkage bias is negligible (i.e. $o(1/\sqrt{n})$).

The next challenge is to construct robust confidence intervals for my LASSO type two-stage estimator. As pointed out in [67], it is impossible to construct uniformly valid confidence intervals for estimates based on model selection. Consistent model selection by LASSO is only guaranteed when all non-zero coefficients are large enough to be distinguished from zero in a finite sample (i.e. usually called the "beta-min" condition). LASSO may fail to select regressors with very small coefficients, resulting in omitted variable bias in the post LASSO inference.

I propose a bias correction for my two-stage estimator following the recent LASSO inference literature [see 10, 89]. The idea is to correct the first order bias and make the estimators independent from the model selection. Heuristically, shrinkage bias due to the l_1 penalty in LASSO can be expressed as a function of the LASSO estimators. Normality can still be achieved after adding back this bias. I show that this strategy also works in a two-stage estimation process. I derive the asymptotic normality for my "de-sparse" two-stage LASSO estimator and conduct robust inference including confidence intervals.

My model can be extended to allow for more flexible network structures.

One real world scenario is a network which consists of multiple cliques. Each clique has its local leaders, who only influence individuals within their own cliques but have no influence on individuals outside their cliques. One identification difficulty in this setting is that the number of leaders increases with the number of cliques. Hence, the sparsity assumption can potentially be violated.

To solve the problem in this scenario, I modify my model by bringing back the classical SAR model. I assume that there are both local leaders and global leaders in the network. In contrast to local leaders, global leaders can influence individuals across different cliques. I assume global leaders are sparse and show that identification can be achieved for this modified model. The endogenous effects of local leaders will be captured by the classical SAR model, which becomes an average endogenous effects in the network. The endogenous effects of global leaders, whose influence remains individual-specific, can be identified in the same way as it was in the previous model. If there is no global leader in the network, the model is effectively just the standard SAR model.

Another real world scenario is the existence of multiple types of connections among individuals. For example, connections among individuals can be classified as social (e.g. friendship, kinship) or economic (e.g. lending, employment). It is also important to identify which networks are more efficient at transmitting the endogenous effects.

I model different types of connections as multiple networks. I propose the use of square-root sparse group LASSO to estimate a heterogeneous endogenous effects model with multiple networks. The standard sparse group LASSO penalizes both the l_1 norm and the l_2 norm for each coefficient in each type of connection. I modify the sparse group LASSO by taking the square-root of

the mean square error and thus make the estimation process pivotal. I derive the convergence rate and prove the consistency of selection. To the best of my knowledge, my paper is the first to show statistical properties for square-root sparse group LASSO.

I provide simulation evidence for networks of different sizes and different generating algorithms. The empirical coverage of my proposed estimators is close to the nominal level in all scenarios. Similar results are also found in models with multiple networks and with cliques.

I apply my method to study villagers' decisions to participate in microfinance programs in rural areas of India as in [9]. Among different social and economic networks, my method shows that some networks such as "visit go-come" and "borrow money", are much more effective at influencing villagers' decisions than other networks such as "temple company" and "medical help". I further show that individuals in certain careers such as agricultural workers, Anganwadi teachers and small business owners are more likely to influence villagers.

My proposed methodology can be applied to detect influential individuals in empirical work when there are both leaders and followers. It is important to identify such individuals because we can then study why certain people are more influential than others. On the one hand, we can examine individuals exogenous characteristics and see if any of them contribute to an individual's influence. On the other hand, we can study how the position of an individual within a network may impact her influence by further introducing network formation into the model.

Being able to identify influential individuals could also lead to more effective policy outcomes. If individuals with certain characteristics are found to be more influential than others, policy makers could potentially implement policies solely targeting influential individuals rather than the entire population. Since more resources are directed to the small group of highly influential individuals, one would expect much more effective policies. For example, online opinion leaders have influence on what people tweet and share on the Internet. In an election, instead of advertising on television and trying to influence every voter, a candidate could invest in these online opinion leaders and let them influence the public in a more efficient way. This technique could also work in employment contexts. Union leaders are often those workers who have the strongest influence on their fellow workers' opinions. Instead of reading through complaints from every worker, employers could identify those union leaders and make sure their complaints were addressed to prevent any ongoing strike. When studying peer effects in smoking behavior, my method can identity a group of teenagers who have a strong influence on their peers. A policy can target this group of students and encourage them to quit smoking.

1.1.1 Literature Review

This paper brings together literature on spatial autoregression model, LASSO and networks.

SARs:

SARs have been used widely applied in empirical studies. For instance, they have been used to study peer effects in labor productivity [see 76, 47, 8], smok-

ing behavior among teenagers [see 59, 27, 79], educational achievements among different student groups [see 82, 80], systemic risk in finance [see 17, 34], and the adoption of new agricultural technologies [see 29, 30]. My paper proposes a novel extension of standard SAR models that could be used to identify influential individuals in any given network. My methodology for estimating such a model could easily be adopted in existing empirical SAR analyses to identify influential individuals who influence their peers productivity, smoking decisions, or financial holdings.

More specifically, my model extends existing SARs literature by introducing *heterogeneous* endogenous effects. Until very recently, SARs always assume a constant rate of dependence for endogenous effects across different individuals (see [28], the first monograph on the topic, and the later studies, [88, 5, 32, 65, 61, 53]). Recent developments in social interaction literature incorporate individual characteristics into SARs, essentially modeling the heterogeneity through a linear combination of exogenous effects [see 74, 19]. In contrast, my model considers the heterogeneity in the endogenous effects. Heterogeneous endogenous effects can be identified from individuals' outcomes instead of being pre-specified through individuals' characteristics. To my knowledge, my proposed model is the first to capture the direct impact of an individuals neighbors' decisions on her own decision.

To estimate the heterogeneous endogenous effects in my model, I propose a methodology that is different from standard SARs literature. In classic SAR models, there is only one endogenous variable and hence it is sufficient to identify the model through only one instrument. In my model, the number of potentially endogenous variables increases as the number of observations increases.

As a result, I propose a set of instruments that contain the same number of instruments as the total number of individuals. Moreover, each instrument is different from the standard SARs instrument as in [58], [62], [63] and [64].

This paper also contributes to literature that models multiple networks through SARs. In standard SARs, multiple networks are modeled as higher order spatial lags [see 65]. Even though different networks are assumed to have different constant rates for endogenous effects in these models each individual in a given network faces the same constant rate. In contrast, my model allows for the a more realistic scenario where each individual has her own specific endogenous effects in each network. Moreover, my methodology allows some networks to be classified as completely irrelevant to decision-making *ex ante* and these networks can be consistently identified.

LASSO:

My paper extends LASSO literature by deriving statistical bounds and consistency of selection for the square-root *sparse group* LASSO estimator. This estimator builds on the group LASSO, square-root LASSO, square-root group LASSO, and sparse group LASSO. [13] introduced the square-root LASSO, which does not require a pre-estimation of an unknown standard deviation σ. [91] proposes the group LASSO, in which explanatory variables are represented by different groups. The group LASSO assumes that sparsity exists only among groups, i.e. some groups of variables are relevant while other groups are not. [84] proposes the sparse group LASSO, which further allows sparsity within each group, i.e. some regressors within the relevant groups can also be irrelevant. [22] derives statistical properties for the square-root group LASSO, which combines group LASSO and square-root LASSO. When estimating a heterogeneous endogenous

effects model with multiple networks, I provide proof for both statistical bounds and consistency of selection for the square-root *sparse group* LASSO estimator. To the best of my knowledge, this paper is the first to show asymptotic statistical properties for the square-root *sparse group* LASSO estimator.

This paper also contributes to the growing literature on endogenous regressors in LASSO estimators. For instance, [11] proposes the double selection mechanism to study confounded treatment effects. [41] proposes a GMM type estimator to deal with many endogenous regressors. [43] proposes a Self Tuning Instrumental Variables (STIV) estimator. The paper that is closest to mine is [94], which studies the statistical properties of two-stage least square procedure with high-dimensional endogenous regressors. She studied a case when there exists p endogenous regressors. For each regressor j, she assumed that one can find d_j instruments. Both p and d_j may grow as n increases. I consider a case that is tailored to my SAR model. There are n endogenous regressors and each regressor shares the same n instruments. I show that a modified "de-sparse" LASSO estimator can be constructed for my estimator in a manner similar to [92], [20], [89] and [94]. I derive its asymptotic distributions and show how to perform inference.

Network:

My paper shares similar microfoundations with SARs as discussed in [16], where the individual utility function can be written as a linear summation of the private and social components. The private component is a quadratic loss function on individual's efforts. The social components depend on the network structure as well as the efforts of one's neighbors. While the marginal rate of substitution between the private and social components of utility is assumed

fixed in SARs, I assume this rate is individual-specific and depends on one's neighbors. My paper applies and extends LASSO approaches to deal with a high-dimensional problem in networks. The total number of possible edges in a network is n^2, however, the social interaction networks we often observe are far more sparse. This is an ideal setting where LASSO could be applied. [72] studies the heterogeneous exogenous effects in a network using LASSO. [33] explore the use of LASSO to recover network structures. Both these two papers consider panel data and rely on repeated observations of the same network to identify their models. My model considers cross-sectional data. To identify an individual's endogenous effects, I rely on the variations in her neighbors' outcome.

My paper also relates to the literature on identifying the key players in the network following [7], [23], and [52]. Under the framework of SARs, every individual is assumed to have the same endogenous effects. As a result, individuals who are well-connected in the network (with high centrality measure) become the key players in the network. However, this is not necessarily the case in my model, as well connected individuals can have zero endogenous effects on her neighbors. Indeed, as shown in the empirical application, well connected villagers such as tailors, hotel workers, veterans, and barbers are not influential in other villagers' decisions to join the micro-finance program.

The rest of this paper is organized as follows: in Section 2, I introduce the model; in Section 3, I discuss identification assumptions; in Section 4, I design estimation procedures; in Section 5, I derive consistency and asymptotic properties; in Section 6, I show finite sample performance using Monte Carlo simulations; in Section 7, I apply my proposed model to study influential individuals

and effective networks in promoting micro-finance programs in rural India; and in Section 8, I conclude.

1.2 Models

In this section, I first lay out the benchmark endogenous effects model and introduce the central model of this paper the heterogeneous endogenous effects model. Then I discuss two extensions of the heterogeneous endogenous effects model: a model for networks consisting of multiple cliques and a model for multiple networks. Finally, I provide two examples and illustrate how my model fits into these settings.

1.2.1 Benchmark Endogenous Effects Model

In this paper, I first introduce the standard spatial autoregression model (SAR) as the benchmark endogenous effects model. Let n denote the total number of observed individuals in a network. The outcome of individual i is denoted as d_i and is the variable of interest. Here d_i can represent any outcome associated with individual i, such as whether to smoke, whether to join a program, or whether to tweet a message from a friend. It is assumed that the outcome of each neighbor of individual i impacts her outcome homogeneously through a constant rate λ_0:

$$d_i = \lambda_0 \sum_{j \in N_i} d_j + x_i \beta_0 + \epsilon_i, \tag{1.1}$$

where the set N_i is defined as individual i's neighbors. The matrix form of this model is expressed as follows:

$$D_n = \lambda_0 M_n D_n + X_n \beta_0 + \epsilon_n, \tag{1.2}$$

where $D_n = (d_1, d_2, \cdots, d_n)'$ is the n-dimensional vector of observable outcomes. The n by k matrix X_n represents the observable exogenous characteristics of individuals. When ϵ_n is specified as an n-dimensional vector of independent and identically distributed disturbances with zero mean and a constant variance σ^2, equation (1.2) is also called a mixed regression model.

The spatial weight matrix M_n is of size n by n, where the (i, j)th entry represents the connection between individual i and individual j. In empirical studies, the spatial weight matrix is often replaced by the adjacency matrix [see 3, 2, 9]: the (i, j)-th entry of the matrix M_n takes value 1 if individual i and individual j are connected and takes value 0 otherwise.

In this model, endogenous effects [see 74] or network effects [see 19] are captured by the scalar λ_0. An implicit assumption in equation (1.2) is that λ_0, the rate of endogenous effects, is identical across all individuals in the network. Every individual affects her neighbors at this same rate λ_0 no matter who she is, how many neighbors she has and where is she in the network. This limitation has been noted in various studies [see 3, 33]. I relax this assumption by proposing a more flexible model that allows individual-specific endogenous effects as discussed below.

1.2.2 Heterogeneous Endogenous Effects Model

I propose the following model to allow for heterogeneous endogenous effects:

$$d_i = \sum_{j \in N_i} d_j \eta_j + x_i \beta + \epsilon_i \tag{1.3}$$

where N_i represents the set of individual i's neighbors and η_j represents the endogenous effects of individual j on the outcome of all her neighbors $i \in N_j$. the model can be rewritten in matrix form as:

$$D_n = \Big(M_n \circ D_n\Big)\eta_0 + X_n \beta_0 + \epsilon_n, \tag{1.4}$$

where $\eta_0 = (\eta_1, \eta_2, \cdots, \eta_n)'$ is a vector of parameter of size n by 1. The ith entry in η_0 represents the endogenous effects of individual i on her neighbors. This model allows for individual heterogeneity to interact with endogenous effects so that every individual is allowed to have her own coefficient η_i. My model allows some $\eta_j = 0$. In other word, there are individuals that have no endogenous effects on their neighbors. I define those individuals with $\eta_j \neq 0$ as influential.

The operator $\circ$ is defined between a n by n matrix M_n and a n by 1 vector D_n as

$$M_n \circ D_n = M_n \cdot \ \text{diag}(D_n) = C,$$

where $diag(\cdot)$ is the diagonalization operator and $C_{i,j} = M_{i,j} d_j$.

Note that in contrast to fixed rate λ_0 specified in equation (1.2), even though each neighbor of individual j is assumed to receive the same influence $d_j \eta_j$ from her[1], each individual is allowed to influence her neighbors at her own rate η_j.

Similar to equation (1.2), equation (1.4) can be derived from a bayesian Nash

[1]Further relaxation of the model considering different individual j's influence on each of her neighbors requires panel data.

Equilibrium. Let (x_i, ϵ_i) denotes an individual's type, where x_i is publicly observed characteristics and ϵ_i is private characteristics only observable by i. Individual i's utility depends on her own action and characteristics as well as her neighbors' actions. Individual i chooses action d_i to maximize the following utility:

$$U_i(d_i, d_{-i}) = (x_i\beta + \epsilon_i)d_i - \frac{1}{2}d_i^2 + \sum_{j \in N_i} d_j d_i \eta_j$$

The first order condition yields equation (1.4)

1.2.3 Examples

To help readers conceptualize the heterogeneous endogenous effects model, here I apply the model to two specific contexts one invloving labor productivity and the other online opinion leaders.

- **Peer Effects in Labor Productivity**

  Understanding the mechanism and magnitude of the dependence of labor productivity on coworkers is an important question for economists and policy makers. As found in [76], workers respond more to the presence of coworkers with whom they frequently interact. In this case, the influence level of each individual to hers coworkers is not necessarily the same. Equation (1.4) can be used to incorporate such differences.

  $$y_i = \sum_{j \in N_i} y_j \eta_j + x_i\beta + \epsilon_i,$$

  where y_i is individual i's productivity, x_i represents individual i's characteristics (education levels, ages, etc) and N_i is the set of coworkers that works directly with i. η_j represents the size of influence of coworker j –

all else being equal, the additional effect on individual i's productivity if individual j becomes her coworker

Note that if we restrict the parameters η_j to be the same across different workers, then we are back to the classical SAR setting as laid out in equation (1.2). Thus, $\lambda = \frac{1}{n}\sum_{j=1}^{n}\eta_j$ can be interpreted as the averaged spillover effects in the canonical sense.

Define

$$\lambda^* = \frac{1}{\sum \mathbf{1}_{\eta_j \neq 0}} \sum_{j=1}^{n} \eta_j \mathbf{1}_{\eta_j \neq 0}$$

as the averaged endogenous effects for influential workers. λ^* does not include non-influential individuals in the calculation. It is a more precise measure of endogenous effects compared with λ from equation (1.2).

- **Online Opinion Leaders**

  A decision can represent whether to "tweet" a news story seen online. When individuals make such decisions, they are often influenced by several online opinion leaders – whether those people "tweet" the news or not. There are also many types of online opinion leaders, including political figures and some are celebrities. For certain types of news, some opinion leaders may be very influential while the rest may have no influence on the public. Opinion leaders may also influence each other when deciding whether to "tweet" the news or not. Assume a binary decision $(0, 1)$ is made from a bayesian Nash Equilibrium, such that

$$d_i^* = \sum_{j \in N_i} d_j^* \eta_j + x_i \beta + \epsilon_i,$$

  where d_i^* is the probability of individual i playing action 1, and $\sum_{j \in N_j} d_j^* \eta_j$ is the expected endogenous effects from i's neighbors N_i . X_i is the individual i's characteristics such as political views, age, career, etc.

Similarly, we can define $\lambda = \frac{1}{n}\sum_{j=1}^{n}\eta_j$ as the averaged endogenous effects. Since the number of opinion leaders is very small compared with total online users, λ can be very close to 0. A more precise measure would be

$$\lambda^* = \frac{1}{\sum \mathbf{1}_{\eta_j \neq 0}} \sum_{j=1}^{n} \eta_j \mathbf{1}_{\eta_j \neq 0}$$

λ^* will be the average endogenous effects for online opinion leaders. On the other hand, it is also important to identify the set:

$$S = \{j : \eta_j \neq 0\}$$

as truly influential opinion leaders. If a similar type of news story needs to be spread the next time, contacting those leaders and obtaining their endorsement would be a good starting strategy.

1.2.4 Heterogeneous Endogenous Effects Model with Cliques

I propose an extension to my heterogeneous endogenous effects model which could address such challenges. Consider a network composed of many cliques (small groups of connected individuals). Each clique has its local leader who only influences individuals within her own clique. Figure 1.1 provides an example of such a network structure. Note that in Figure 1.1, node S_2, S_3 and S_4 represent local leaders who only influence individuals within their own cliques. On the contrary, node S_1 represents a global leader who can influence individuals across different cliques. For example, one can think about the local leaders S_2, S_3 and S_4 as local news channels while S_1 is the national news channel. I assume that all local news will influence the public at a small but similar rate while different national channels can have different effects on their audience.

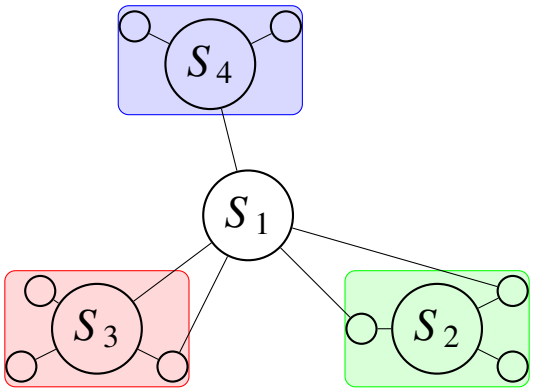


Figure 1.1: Local Leader

In the above network structure, if the number of local leaders is increasing with the number of cliques but the number of individuals in each clique stays fixed, it is impossible to identify the individual-specific influence of all those local leaders. To address this problem, I assume a homogeneous effect γ_0 among all individuals. This rate will capture all influence from local leaders. However, I allow global leaders to heterogeneously influence their neighbors at rates that differ from γ_0 and show that γ_0 and the heterogeneous effects can be consistently estimated.

More specifically, I consider the following model:

$$d_i = \sum_{j \in N_i} d_j \eta_j + \gamma_0 \sum_{j \in N_i} d_j + x_i \beta_0 + \epsilon_i, \tag{1.5}$$

which be represented in matrix form as:

$$D_n = \Big(M_n \circ D_n\Big)\eta_0 + M_n D_n \gamma_0 + X_n \beta_0 + \epsilon_n, \tag{1.6}$$

where $\eta_0' = (\eta_1, \eta_2, \cdots, \eta_n)'$. The new term $\gamma_0 \sum_{j \in N_i} d_j$ captures influence from the local level. Note that this is the same term as the spatial lag in the benchmark SAR model. The vector η_0 captures the heterogeneous endogenous effects of global leaders.

If no global leader exists, i.e. $\eta_j = 0, \forall j$, the model collapses back to the classical SAR model as in equation (1.2). If there is no local level influence, i.e. $\gamma_0 = 0$, then the model coincides with the heterogeneous endogenous effects model in section 2.2.

1.2.5 Heterogeneous Endogenous Effects Model with Multiple Networks

In reality, individuals are often connected with each other through more than one type of network. For example, ones colleague (connection in an employment network) could also be her friend (connection in a friendship network), and ones uncle (connection in a relative network) could also be the person she lends money to (connection in a borrowing/lending network). In such scenarios, an individuals outcome could potentially be influenced by the outcomes of her neighbors from more than one type of network.

To capture different types of connections among the same set of individuals, we can incorporate multiple networks in my heterogeneous endogenous model. More specifically, a separate adjacency matrix can be constructed for each type of network. For instance, the (i, j)-th entry of the adjacency matrix representing friendship takes value 1 if individual i and individual j are friends and takes value 0 otherwise; that representing the borrowing/lending network takes value 1 if individual i and individual j lend money to each other and takes value 0 otherwise.

Let q be the total number of different types of network. Define M_n^l as the

adjacency matrix for the lth network. The heterogeneous endogenous effects model with multiple networks is defined as

$$d_i = \sum_{l=1}^{q} \sum_{k \in N_i} d_k^l \eta_k^l + x_i \beta_0 + \epsilon_i \tag{1.7}$$

Note that in this model, different network could potentially bear different endogenous effects for the same individual. In equation (1.7), coefficient η_k^l represents the rate of endogenous effect of individual k through network l. As a result, we have $nq + k$ coefficients for endogenous effects. In addition, I assume endogenous effects from different types of networks are linearly additive. The model can also be rewritten in matrix form as:

$$D_n = \sum_{l=1}^{q} \left(M_n^l \circ D_n \right) \eta_0^l + X_n \beta_0 + \epsilon_n, \tag{1.8}$$

where M_n^l is the adjacency matrix for network l. $\eta^l = (\eta_1^l, \eta_2^l, \cdots, \eta_n^l)'$ is an n by 1 vector for $l = 1, 2, \cdots, q$. Define a network l as efficient network if $\eta_i^l \neq 0$ for at least one individual $i = 1, 2, \cdots, n$.

1.3 Identification

In this section, I discuss the conditions under which the heterogeneous endogenous effects model is identified and the extensions of this model. My assumptions combine both standard SARs type assumptions and LASSO type assumptions. SARs type assumptions ensure the existence of valid instruments to identify the model. LASSO type assumptions guarantee consistent model selection and estimation using a LASSO estimator. In what follows, I will first present the assumptions needed for a standard heterogeneous endogenous effects model to be identified. Then I will discuss identification assumptions for two model extensions laid out in the previous section – one heterogeneous endogenous effects

model for networks consisting of multiple cliques and one with multiple types of networks.

Before discussing identification assumptions for the heterogeneous endogenous effects model, lets first recall the benchmark SAR model:

$$D_n = \lambda_0 M_n D_n + X_n \beta_0 + \epsilon_n, \tag{1.9}$$

Note that by rearranging the above equation, we can express endogenous variable $M_n D_n$ solely as a function of X_n and M_n, since:

$$D_n = J_n^{-1} X_n \beta_0 + J_n^{-1} \epsilon_n$$

where I_n is the n by n identity matrix and $J_n = I_n - \lambda_0 M_n$. It is straightforward that $J_n^{-1} X_n$ can serve as valid instruments for $M_n D_n$. As a result, the identification and estimation of equation (1.9) can be achieved through either 2SLS or GMM as proposed in papers such as [58], [57] [62], [63], and [64].

As will be explained in detail in subsequent sections, to estimate the individual specific effects in the heterogeneous endogenous effects model, I derive a set of instruments in a similar way by solving D_n as a function of exogenous variables and an adjacency matrix. The assumptions listed below essentially guarantee the existence and consistency of the 2SLS estimates.

1.3.1 Identification Assumptions for the Heterogeneous Endogenous Effects Model

Recall that the heterogeneous endogenous effects model is specified as:

$$D_n = \Big(M_n \circ D_n\Big)\eta_0 + X_n \beta_0 + \epsilon_n,$$

First note that without additional restrictions, this model could not be point identified through canonical method as the number of parameters $n+k$ is greater than the number of observations n. To achieve identification, the key assumption that I maintain is that only a small number of individuals in the network are influential (i.e. $\eta_j \neq 0$).

Assumption A. *Let $S_n \subset \{1, 2, \cdots, n\}$ denote the set of influential individuals (i.e. $\eta_j \neq 0$). Let $s_n = |S_n|$ be the number of elements in S_n.*

$$s_n = o\left(\frac{\sqrt{n}}{\log n}\right), \quad \text{as } n \to \infty$$

Assumption 1 is usually referred to as "sparsity" assumption. The assumption that most individuals in a network are not influential is plausible under many circumstances. For example, opinion leaders on social media only constitute a very small fraction of internet users; there are only a couple of "cool" kids at school that might influence their friends' smoking decisions; passionate workers that can boost the productivity of their coworkers are also relatively rare. When many local leaders exist within a network, the sparsity assumption could be violated. I will address this issue in section 3.2 and show that identification can still be achieved with additional assumptions.

Assumption B.

- *There exists an $\eta_{\max} < 1$ such that $\|\eta_0\|_\infty \leq \eta_{max}$*
- *The ϵ_j are i.i.d with 0 mean and variance σ^2*
- *The regressors x_i in X_n are non-stochastic and uniformly bounded for all n. $\lim_{n\to\infty} X_n'X_n/n$ exists and is nonsingular*

Assumption 2 guarantees the invertibility of $\left(I_n - M_n \circ \eta_0\right)$. The restriction on η_0 excludes the unit root process and ensures the uniqueness of equilibrium. The assumptions on the error term and the assumption that X_n is a fixed design matrix are the same as those imposed in the mixed regression model[2] [see 62]. I focus on the case where X_n is an n by 1 vector and study identification as in [19]. It is straightforward to generalize the algebra when X_n is n by k. More instruments can be constructed in this scenario.

To proceed, recall the definition of the operator "$\circ$" as $M_n \circ D_n = M_n \cdot \text{diag}(D_n)$, where $\text{diag}(\cdot)$ is the diagonalization operator. Note the following property of the "$\circ$":

$$\left(M_n \circ D_n\right)\eta_0 = \left(M_n \circ \eta_0\right)D_n$$

If the invertibility of $\left(I_n - M_n \circ \eta_0\right)$ is guaranteed, then

$$D_n = \left(M_n \circ D_n\right)\eta_0 + X_n\beta_0 + \epsilon_n \Leftrightarrow D_n = \sum_{i=0}^{\infty}\left(M_n \circ \eta_0\right)^i(X_n\beta_0 + \epsilon_n) \tag{1.10}$$

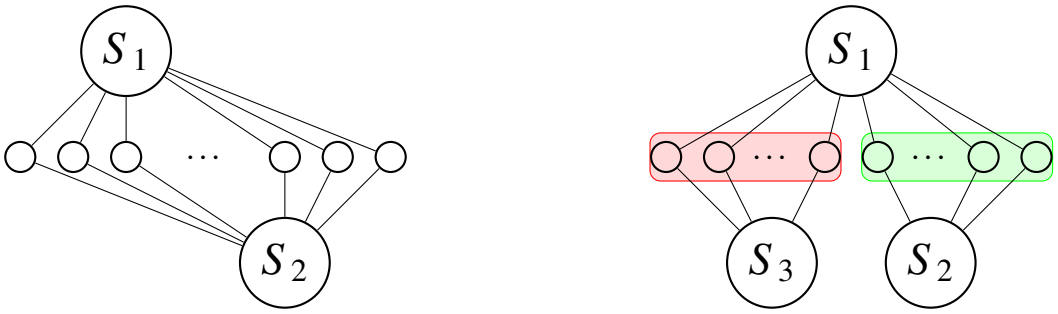
This is formally shown in Appendix B.

Since $\left(M_n \circ D_n\right)\eta_0$ is correlated with ϵ_n and η_0 is sparse (i.e. having at most s_n non-zero elements), we need at least s_n instruments to deal with the endogeneity in the model. Using equation (1.10), we can express the expectation of D_n as follows:

$$E(D_n) = X_n\beta_0 + \left(M_n \circ X_n\right)(\beta\eta_0) + \sum_{i=2}^{\infty}\left(M_n \circ \eta_0\right)^i\beta_0 X_n, \tag{1.11}$$

Let $(\cdot)_S$ denote the operator such that $(M_n)_S$ is a sub matrix of M_n with its columns restricted to columns corresponding to the elements of S. The first

[2]The assumption on error terms exclude exogenous effects and correlated effects from my model. An identification problem similar to the "reflection problem" arises when including exogenous effects. More instruments need to be constructed, which requires better data. These are interesting directions for future research.



(1) (2)

Figure 1.2: Examples of networks which violate assumption 3

and second terms of equation (1.11) suggest that X_n and $(M_n \circ X_n)_S$ can serve as valid instruments to point identify β_0 and η_0.

Assumption C. $[X_n, (M_n \circ X_n)_S]$ *is full rank.*

Assumption 3 is the key assumption that leads to identification. The linear independence among $(M_n \circ X_n)_S$ requires the assumption that any two influential individuals may not necessarily connect with identical neighbors. Moreover, assumption 3 also requires that neighbors of an influential individual cannot be a linear combination of neighbors of several other influential individuals, which rules out network structures as depicted in Figure 1.2:

In other words, as long as each influential individual has a neighbor that is not connected with any other influential individuals, assumption 3 is satisfied. One can think of the identification here as estimating fixed effects from influential individuals. Collinearity arises when the fixed effects of two influential individuals are imposed on exactly the same observations. As shown in Figure

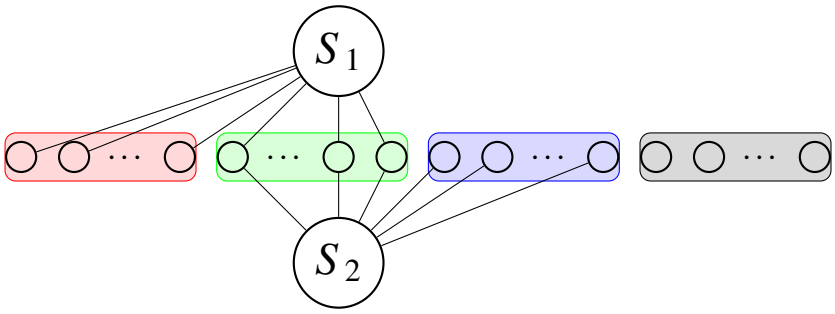


Figure 1.3: Fixed Effects

1.3, the influence of S_1 can be identified by comparing red and yellow groups, while the influence of S_2 can be identified by comparing blue and black groups. Or the influence of S_1 can be identified by comparing green and blue groups, while the influence of S_2 can be identified by comparing red and green groups.

Further, as shown in Appendix B, one can rewrite equation (1.11) as:

$$E(D_n) = X_n\beta_0 + \Big(M_n \circ X_n\Big)\tilde{\eta}, \tag{1.12}$$

where $\tilde{\eta}_j = \eta_j f(\beta_0, X_n, M_n)$ for some function f depends on β_0, X_n, and M_n. Note that $\tilde{\eta}_j = 0$ as long as $\eta_j = 0$. As a result, the sparsity assumption is also satisfied in equation (1.12), and I can thus estimate equation (1.12) as the first stage in using a LASSO type estimator.

At this point, if the truly influential individuals set S_n were available to us, we would be able to estimate the model using 2SLS method or GMM. However, in most cases, S_n is not known beforehand. I propose to use a LASSO type estimator to both recover the set of influential individuals and estimate the model. For LASSO to achieve correct recovery, I need the following assumptions:

Assumption D.

(Irrepresentable Condition) *There exists* $N \in \mathbb{N}$: $\forall n \geq N$, *there is a* $\vartheta \in (0, 1)$ *such*

that

$$P\left(\left\|diag((\hat{D}_n)_{S^c})\Sigma_n diag((\hat{D}_n)_S)^{-1} sign(\eta_0)\right\|_\infty \leq \vartheta\right) = 1;$$

where

$$\Sigma_n = (M_n)'_{S^c}(M_n)_S \left((M_n)'_S(M_n)_S\right)^{-1},$$

(Beta Min Condition) *There exists* $N \in \mathbb{N}$*:* $\forall n \geq N$*, there is a* $m > 0$ *such that*

$$\min(|\eta_0|)_S \geq m/\sqrt{n},$$

Here $(M_n)_S$ represents the sub-matrix of M_n given by the columns corresponding to influential individuals. Similarly, $(M_n)_{S^c}$ represents the sub-matrix of M_n given by the columns corresponding to non-influential individuals.

Assumption 4 is required for the LASSO estimator to achieve a consistent selection for the set S_n in the second stage. The Irrepresentable Condition imposes restrictions on non-influential individuals such that the neighbors of a non-influential individual will not be exactly the same as those of any influential individual. This is because when two individuals connect with exactly the same neighbors, we cannot distinguish which individual is the true source of influence. This assumption rules out identification in complete networks (i.e. all individuals are connected). The Beta Min Condition requires the magnitude of the endogenous effects to be sufficiently strong in order to be detected by LASSO. As shown in [93], the Irrepresentable Condition together with the Beta Min Condition are necessary and sufficient conditions for LASSO to achieve consistent model selection. If consistent selection is not required, these two conditions can also be relaxed to weaker conditions (such as the compatibility condition as in [21]). As shown in [89], with the compatibility condition, inference on the de-sparse coefficients as discussed in the next section is still valid.

Assumption E.

(Maximum Neighbors Condition) *There exists* $N \in \mathbb{N}$*:* $\forall n \geq N$*,*

$$\|M_n' \mathbf{1}_n\|_\infty = O([\log n]^\epsilon), \quad \epsilon \in (0, 1]$$

(Variance Condition)

$$\frac{1}{n} M_n' W_n (I - M_n \circ \eta_0)^{-1} (I - M_n \circ \eta_0)^{-1'} W_n M_n \to \Omega,$$

where $W_n = \left(I - X_n (X_n' X_n)^- X_n'\right)$

The Maximum Neighbors Condition requires the network structure (edges) to be sparse. More specifically, it requires that the number of direct neighbors not increase faster than $O(\log n)$ when the number of influential individuals increases at speed $o\left(\frac{\sqrt{n}}{\log n}\right)$. This rate can be improved when the number of influential individuals is fixed. The Maximum Neighbors Condition is an asymptotic bound on the number of neighbors for each individual as the network increases. This condition is required to prevent shrinkage bias carried from the first stage LASSO estimation from growing faster than $o(1/\sqrt{n})$ in the second stage.

The Variance Condition requires the variance-covariance matrix to converge to a limit. In classical SARs, the spatial weight matrix is assumed to be uniformly row sum bounded. This assumption implies the Variance Condition but imposes restrictions on the network structure. Each individual may only connect with a finite number of neighbors. In my case, the identification of an influential individual comes from the difference in responses between neighbors that solely connect with her and individuals who connect with no influential individuals. For example, consider two groups of individuals that have the same characteristic X where one group all connects with individual j and the other

does not. If the mean response of the two groups is significantly different, we can conclude that j is influential. To identify the influence of individual j as a fixed effect, the number of individuals affected by individual j must grow as the sample size increases. As a result, the row sum for influential individuals cannot be bounded by a fixed number.

The heterogeneous endogenous effects model is identified under assumptions 1-5 as a linear system with a unique solution. I discuss the identification of my model with cliques and with multiple networks in the following two sections.

1.3.2 Identification Assumptions with Cliques

Recall the heterogeneous endogenous effects model with cliques, represented as follows:

$$D_n = \Big(M_n \circ D_n\Big)\eta_0 + M_n D_n \gamma_0 + X_n \beta_0 + \epsilon_n$$

Define global leaders as those influential individuals who influence multiple cliques and whose neighborhoods increase as n increases. Define local leaders as influential individuals who are not global leaders.

Assumption 1′. *Among n individuals in the network, let $S_n \subset \{1, 2, \cdots, n\}$ be the set of global leaders. Let $s_n = |S_n|$ be the number of elements in S_n. Assume:*

$$s_n = o\left(\frac{\sqrt{n}}{\log n}\right), \quad \text{as } n \to \infty$$

Assumption 1′ only requires the number of global leaders to be sparse. My model does not impose any restriction on the number of local leaders. As a re-

sult, it does not rule out situations where everyone is (locally) influential. Local leaders' influence will be captured by the γ_0, coefficient of classical spatial lag.

To ensure invertibility of the matrix $\left(I_n - M_n \circ \eta_0 - M_n\gamma_0\right)$, I modify the first part of assumption 2 as:

Assumption 2'. *There exists an* $\eta_{\max} < 1$ *such that* $||\eta_0 + \gamma_0||_\infty \leq \eta_{max}$

Similar to assumption 2, this assumption excludes unit root processes. Since there exists a local level influence γ_0 in the network, global level influence η_0 needs to be further bounded above by 1. As a result, equation (1.6) can be transformed into the following:

$$E(D_n) = X_n\beta_0 + \left(M_n \circ X_n\right)(\beta_0\eta_0) + M_nX_n(\beta_0\gamma) + \sum_{i=2}^{\infty}\left(M_n \circ \eta_0 + \gamma M_n\right)^i\beta_0X_n$$

Equation (1.6) introduces one more coefficient γ_0 compared with equation (1.4). As a result, assumption 3 is modified to include an extra instrument M_nX_n, which is also the classic instrument used in equation (1.2):

Assumption 3'. $[X_n, (M_n \circ X_n)_S, M_nX_n]$ *is full rank.*

Assumption 3' is similar to assumption 3 and requires the additional instrument M_nX_n to be linearly independent with $[X_n, (M_n \circ X_n)_S]$. The remaining assumptions 4 and 5 are unchanged.

1.3.3 Identification Assumptions with Multiple Networks

Recall the heterogeneous endogenous effects model with multiple networks, represented as follows:

$$D_n = \sum_{j=1}^{q}\left(M_n^j \circ D_n\right)\eta_0^j + X_n\beta_0 + \epsilon_n$$

First notice that the number of coefficients in this model becomes $nq + k$. The number of observed networks q is also allowed to increase as the number of observations increases. As a result, the sparsity assumption will be imposed on both the influential individuals and the effective networks. I assume that some of the networks are completely irrelevant (i.e. $\eta_0^j = 0$) and that relevant networks are not necessarily passing influence for everyone (i.e. $\eta_0^j \neq 0$ but $\eta_{0,i}^j = 0$ for some i).

Second, to ensure invertibility, for any matrix norm $\|.\|$:

$$\left\| \sum_{j=1}^{q} \left(M_n^j \circ \eta_0^j \right) \right\| \leq \sum_{j=1}^{q} \left\| \left(M_n^j \circ \eta_0^j \right) \right\| \leq \sum_{j=1}^{q} \|\eta_0^j\|_\infty \left\| \left(M_n^j \right) \right\|$$

Because M_n^j is the adjacency matrix such that each entry is 0 or 1, $\sum_{j=1}^{q} \|\eta_0^j\|_\infty < 1$ guarantees the invertibility of $I - \sum_{j=1}^{q} \left(M_n^j \circ \eta_0^j \right)$.

Third, I require $\left[X_n, \left(M_n^1 \circ X_n \right)_S, \left(M_n^2 \circ X_n \right)_S, \cdots, \left(M_n^q \circ X_n \right)_S \right]$ to be full rank. Compared with the standard model, this assumption requires the independence condition to hold across different networks. Again, we cannot identify the source of influence if two influential individuals connect to the same neighbors. Fourth, I assume conditions that guarantee a consistent selection of square-root sparse group LASSO. And, finally, the Maximum Neighbor Condition needs to be satisfied in all q adjacency matrices. Since the five conditions for multiple networks are very similar to assumption 1-5, I list them formally in the appendix as assumption 1*-5*.

1.4 Estimation

I propose an estimator similar to the two-stage least square method but use LASSO in both stages. The estimator proposed here is differs from the "double selection" estimator proposed in [11] as I plugin the fitting from the first stage directly to the second stage. It is in the same framework as that proposed in [94]. I call this estimator a two-stage LASSO (2SLSS) estimator. In this section, I define this 2SLSS procedure and propose a bias corrected version of the estimator. I show how this procedure can be extended to estimate my model for networks consisting of multiple cliques and my model for multiple networks.

1.4.1 Two-Stage LASSO Estimator

I propose to estimate equation (1.4) using the following estimator:

Two-Stage LASSO Estimator:

- First Stage:

$$(\tilde{\beta}, \tilde{\eta}) = \arg\min_{\beta,\eta} ||D_n - X_n\beta - \Big(M_n \circ X_n\Big)\eta||_2 + \lambda|\eta|_1$$

  Obtain a LASSO fitting $\hat{D}_n$

$$\hat{D}_n = X_n\tilde{\beta} + \Big(M_n \circ X_n\Big)\tilde{\eta}$$

- Second Stage:

$$(\hat{\beta}, \hat{\eta}) = \arg\min_{\beta,\eta} ||D_n - \Big(M_n \circ \hat{D}_n\Big)\eta - X_n\beta||_2 + \lambda|\eta|_1$$

As shown in section 1.3, $\left(M_n \circ D_n\right)$ is correlated with ϵ_n. Thus equation (1.4), equation (1.6) and equation (1.8) cannot be estimated directly using LASSO or sparse group LASSO. The instruments proposed in section 1.3 are $[X_n, (M_n \circ X_n)_S]$. We do not observe the set S but note that $[X_n, (M_n \circ X_n)]$ is a set of regressors that contains the valid instruments.

The two-stage least square method can be used to address endogeneity in SARs as in [63]. In the first stage, $M_n X_n$ are used as instruments to estimate D_n. In the second stage, $M_n \hat{D}_n$ is used to replace $M_n D_n$ to avoid endogeneity.

Following the same idea, I estimate a first stage using $[X_n, (M_n \circ X_n)]$. Since there are $n+k$ regressors, I use the square-root LASSO to select those instruments in set S. I choose the square-root LASSO over standard LASSO to avoid a pre-estimation of the unknown variance of the error term σ^2. I construct a synthetic $\hat{D}_n$ variable using square-root LASSO estimates. In the second stage, I replace D_n with $\hat{D}_n$ in the regressors and estimate the coefficients $\hat{\eta}$ using the square-root LASSO again.

The statistical properties of two-stage estimators using LASSO have been studied in [94], where she derives bounds for the estimator and proves consistency of model selection in a general setting. [94] studied the over identified case where the number of endogenous regressors goes to infinity while the number of instruments for each regressor also goes to infinity. I studied the just identified case using the instruments proposed in section 3, where the number of endogenous regressors is the same as the number of instruments and both go to infinity.

1.4.2 De-sparse 2SLSS Estimator

The estimator $\hat{\beta}$ and $\hat{\eta}$ suffers from LASSO shrinkage bias. Moreover, post model selection inference conditioning on the selected model $\hat{S}_n = \{i|\hat{\eta} \neq 0\}$ suffers from the omitted variable bias and thus is not uniformly valid. [see 66, 67, 68]. I construct a "de-sparse" estimator under my setting and derive the asymptotic distribution for it. I propose the following de-sparse LASSO estimator:

De-sparse 2SLSS Estimator:

- Define

$$\hat{e} = \hat{\eta} + \hat{\Theta}(M_n \circ \hat{D}_n)'(D_n - X_n\hat{\beta} - (M_n \circ \hat{D}_n)\hat{\eta})/n$$

- Define

$$\hat{b} = \hat{\beta} - (X_n'X_n)^- X_n'(M_n \circ \hat{D}_n)'(D_n - X_n\hat{\beta} - (M_n \circ \hat{D}_n)\hat{\eta})/n$$

$\hat{\beta}$ and $\hat{\eta}$ are estimators from the 2SLSS. $\hat{\Theta}$ is defined by the nodewise regression as in [77]. Nodewise regression explores the correlation between the columns of the design matrix $W_n(M_n \circ \hat{D}_n)$ by regressing each column on all the rest of the columns while penalizing the coefficients. An approximation of the inverse of the matrix $\frac{1}{n}(M_n \circ \hat{D}_n)'W_n(M_n \circ \hat{D}_n)$ can be constructed based on nodewise regression. Further, define $\hat{S}_n = \{i|\hat{\eta} \neq 0\}$, which represents the LASSO selected active set. The estimators $(\hat{e}, \hat{b})$ are adjusted for the LASSO shrinkage bias and are a consistent estimator for β and η. They are similar to the estimators proposed in [89], but are constructed through a two-stage process as well as using square-root LASSO.

The de-sparse LASSO estimator does not depend on the selected active set. Thus, it does not suffer from the non-uniformity problem. Notice that the double selection method proposed in [11] could also be applied to conduct inference on $\hat{\beta}$. [12] shows the first order equivalence of the double selection method and the de-sparse method. On the other hand, the main interest of this paper is the coefficients $\hat{\eta}$. The double selection method does not provide a way to conduct inference for all the coefficients in the model, while the de-sparse LASSO estimator does.

My de-sparse LASSO estimator differs from the one proposed in [94]. Since the instruments are known in my case, I can derive the asymptotics for my estimator explicitly. By considering a sparse network structure (e.g. Maximum Neighbors Condition), I can show that the shrinkage bias from the first stage is negligible ($o(1/\sqrt{n})$). The estimator proposed in [94] adjusts shrinkage bias from both the first and second stages. In order to show consistency, she assumes the convergence of the product between the residual of nodewise regression and the endogenous regressors.

I will defer the proof of consistency for the LASSO selected set $\hat{S}_n$ and consistency and asymptotic distribution for my estimator $(\hat{e}, \hat{b})$ to section 1.5. In the remainder of this subsection, I will define the estimators for the two extended models.

1.4.3 2SLSS with Cliques

To estimate equation (1.6), I propose the following 2SLSS:

Two-Stage LASSO Estimator with Homogenous Effects:

- First Stage:

$$(\tilde{\beta}, \tilde{\gamma}, \tilde{\eta}) = \arg\min_{\beta,\gamma,\eta} ||D_n - X_n\beta - M_nX_n\gamma - \Big(M_n \circ X_n\Big)\eta||_2 + \lambda(|\eta|_1 + |\gamma|)$$

Obtain a LASSO fitting $\hat{D}_n$

$$\hat{D}_n = X_n\tilde{\beta} + M_nX_n\tilde{\gamma} + \Big(M_n \circ X_n\Big)\tilde{\eta}$$

- Second Stage:

$$(\hat{\beta}, \hat{\gamma}, \hat{\eta}) = \arg\min_{\beta,\gamma,\eta} ||D_n - M_n\hat{D}_n\gamma - \Big(M_n \circ \hat{D}_n\Big)\eta - X_n\beta||_2 + \lambda(|\eta|_1 + |\gamma|)$$

The estimator is similar to that for the previous model except that the classical spatial lag M_nX_n is now included in the estimation. In the above estimator, I penalize ηs and γ at the same rate because I have no prior knowledge of these two effects. One can penalize them at a different rate or not penalize γ if one believes that influence from local leaders is more likely than that from global leaders or vice versa. Since γ and ηs are both penalized coefficients, a similar de-sparse LASSO estimator can be constructed for γ:

De-sparse 2SLSS Estimator with Cliques:

- Define

$$\hat{r} = \hat{\gamma} + \hat{\Theta}(M_n\hat{D}_n)'(D_n - X_n\hat{\beta} - M_n\hat{D}_n\tilde{\gamma} - (M_n \circ \hat{D}_n)\hat{\eta})/n$$

Note that $\hat{\Theta}$ should be an approximation for the inverse of the matrix $\frac{1}{n}[M_n\hat{D}_n, (M_n \circ \hat{D}_n)]' W_n [M_n\hat{D}_n, (M_n \circ \hat{D}_n)]$ in this case.

1.4.4 Multiple Networks

When multiple networks exist, each individual will have network-specific endogenous effects. The number of unknown coefficients increases from $n + k$ to $nq + k$ compared with the standard case. These coefficients can also be classified into q different groups based on networks. By applying the sparsity assumption to the relevant networks, we can estimate the model using the square-root sparse group LASSO instead of the square-root LASSO and propose the following estimator. The square-root sparse group LASSO penalizes both the l_1 and l_2 norm in each group. It can identify all the relevant groups under weaker assumptions compared with the square-root LASSO estimator.

Two-Stage LASSO Estimator with Multiple Networks:

- First Stage:

$$(\tilde{\beta}, \tilde{\eta}) = \arg\min_{\beta,\eta} \left\{ \left\| D_n - X_n\beta - \sum_{j=1}^{q} (M_n^j \circ X_n)\eta^j \right\|_2 + \left(\sum_{j=1}^{q} \left(\lambda_1 \|\eta^j\|_2 + \lambda_2 \|\eta^j\|_1 \right) \right) \right\}$$

  Obtain a LASSO fitting $\hat{D}_n$

$$\hat{D}_n = X_n\tilde{\beta} + \sum_{j=1}^{q} (M_n^j \circ X_n)\tilde{\eta}^j$$

- Second Stage:

$$(\hat{\beta}, \hat{\eta}) = \arg\min_{\beta,\eta} \left\{ \left\| D_n - X_n\beta - \sum_{j=1}^{q} (M_n^j \circ \hat{D}_n)\eta^j \right\|_2 + \left(\sum_{j=1}^{q} \left(\lambda_1 \|\eta^j\|_2 + \lambda_2 \|\eta^j\|_1 \right) \right) \right\}$$

The square-root sparse group LASSO introduces two tuning parameters, λ_1 and λ_2, to penalize both the l_1 and the l_2 norm in each network. Similar to the LASSO estimator, the geometric shape of the penalties allows the square-root sparse group LASSO to identify sparsity not only within each network (group) but also among networks (groups). In other words, some networks could be completely irrelevant (i.e. $\eta^j = 0$) and within relevant networks, some individuals can have no influence on their neighbors (i.e. $\eta^j \neq 0$ but $\eta_i^j = 0$ for some i). The sparse group Lasso was first proposed by [84]. They provide an algorithm to solve this problem without deriving any statistical properties. I modify the estimator by taking the square-root of the mean square error term in the minimization problem. Similar to the square-root LASSO proposed in [13], the method becomes pivotal since it does not require a pre-estimation of the stan-

dard deviation σ. I will prove the statistical properties of square-root sparse group LASSO in section 1.5.

The de-sparse LASSO estimator for square-root sparse group LASSO is proposed as follows:

De-sparse 2SLSS Estimator for Square-root Sparse Group LASSO:

- Define

$$\hat{e}_m = \hat{\eta} + \hat{\Theta}_Z \hat{Z}_n'(D_n - \hat{Z}_n\hat{\eta} - X_n\hat{\beta})/n$$

$$\hat{b}_m = \hat{\beta} - (X_n'X_n)^{-1}X_n'\hat{Z}_n\hat{\Theta}_Z X_n'(D_n - \hat{Z}_n\hat{\eta} - X_n\hat{\beta})/n$$

---

where $\hat{Z}_n = \left[(M_n^1 \circ \hat{D}_n), (M_n^2 \circ \hat{D}_n), \cdots (M_n^q \circ \hat{D}_n)\right]$ and $\hat{\Theta}_Z$ is the approximation of the inverse of the matrix $\frac{1}{n}\hat{Z}_n' W_n \hat{Z}_n$.

1.5 Statistical Properties

In this section, I consider the statistical properties for the de-sparse 2SLSS estimators $(\hat{e}, \hat{b}, \hat{S}_n)$ proposed in section 1.4. I show consistency and derive asymptotic normality for my de-sparse estimators. In order to show consistency and asymptotic normality for the de-sparse 2SLSS estimator with multiple networks, I derive the statistical properties for square-root sparse group LASSO, which have not been previously defined in statistics literature.

1.5.1 Consistency

The proof of consistency has two parts. 1) I show that the selected active set converges to the true non-zero parameter set. 2) I show that the de-sparse estimators converge to the true parameters.

Theorem 1. *In heterogeneous endogenous effects model and with assumption 1-5, if* $\lambda \propto \sqrt{\frac{\log n}{n}}$

- $\lim_{n\to\infty} \mathbb{P}(\hat{S}_n = S) = 1$
- $\hat{e} \to \eta_0$
- $\hat{b} \to \beta_0$

The consistency of the LASSO active set $\hat{S}_n$ follows from assumption 4 as is shown in [93]. The consistency of $\hat{e}$ and $\hat{b}$ can be shown by taking the Karush-Kuhn-Tucker conditions of the LASSO minimization problem in the second stage. The shrinkage bias carried from the first stage: $\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (\hat{D}_n - D_n)\Big)\eta_0$ can be shown of order $o(1/\sqrt{n})$. The details of this proof are provided in the appendix.

In the presence of cliques, if γ is penalized, it can be treated as one of the components in η. On the other hand, if it is not penalized, it can be treated as one of the components in β. The consistency follows directly from Theorem 1:

Corollary 1. *In the heterogeneous endogenous effects model with cliques and under assumptions 1'-3', assumptions 4-5, if* $\lambda \propto \sqrt{\frac{\log n}{n}}$

- $\lim_{n\to\infty} \mathbb{P}(\hat{S}_n = S) = 1$
- $\hat{e} \to \eta_0$

- $\hat{r} \to \gamma_0$
- $\hat{b} \to \beta_0$

In the presence of multiple networks, theorem 2 summarizes the consistency results.

Theorem 2. *In the heterogeneous endogenous effects model with multiple networks and under assumptions 1*-5*, if* $\lambda_1 \propto \sqrt{\frac{\log n}{n}}$ *and* $\lambda_2 \propto \sqrt{\frac{\log n}{n}}$

- $\lim_{n\to\infty} \mathbb{P}(\hat{S}_n = S) = 1$
- $\hat{e}_m^j \to \eta_0^j$ *for* $j = 1, \cdots, q$
- $\hat{b}_m \to \beta_0$

The derivation of theorem 2 is similar to that of theorem 1 expect that the square-root LASSO is replaced with the square-root sparse group LASSO. Theorem 1, corollary 1 and theorem 2 establish the consistency for my de-sparse 2SLSS estimators.

1.5.2 Asymptotics

Post inference or inference after model selection are not uniformly valid. Define the set:

$$B(s) = \{\eta \in \mathbb{R}^n | \{j, \eta_j \neq 0\} \leq s\}$$

As shown in [66], [67], and [56]

$$\sup_{\eta_0 \in B(s)} \left| P\left(\frac{\sqrt{n}(\hat{\eta}_j - \eta_0)}{\hat{V}_j} < t \right) - \Phi(t) \right| \nrightarrow 0 \tag{1.13}$$

where $\hat{\eta}_j$ can be any estimator based on a selected model, $\hat{V}_j$ is the associated standard deviation and $\Phi(t)$ is the normal CDF function. When η_j is of order $O(1/\sqrt{n})$, the probability that LASSO fails to select this regressor into the active set can be non-zero. The resulting post model selection estimator will carry the omitted variable bias because of the exclusion of regressor j from the model. Thus, post inference conditioning on the selected model cannot converge to the true parameters uniformly over the models defined by sparsity.

On the other hand, the de-sparse LASSO estimator is uniformly valid since the inference is not conditioned on the selected model [see 89]. I follow the same idea and show that my de-sparse 2SLSS estimators achieve asymptotic normality with square-root LASSO and square-root sparse group LASSO.

Theorem 3. *In the heterogeneous endogenous effects model and under assumption 1-5, if* $\lambda \propto \sqrt{\log n/n}$

$$\sqrt{n}(\hat{e} - \eta_0) = E_1 + \Delta_1,$$
$$\sqrt{n}(\hat{b} - \beta_0) = E_2 + \Delta_2,$$

where

$$E_1 \sim N(0, \sigma^2 \Theta_1 diag(\Gamma)\Omega diag(\Gamma)\Theta_1'),$$
$$E_2 \sim N(0, \sigma^2 \Theta_2 diag(\Gamma)\Omega diag(\Gamma)\Theta_2'),$$

and

$$\|\Delta_1\|_\infty = o_p(1), \quad \|\Delta_2\|_\infty = o_p(1),$$
$$\Gamma = \lim_{n\to\infty}(I - M_n \circ \eta_0)^{-} X_n \beta_0,$$
$$\Theta_1 = \lim_{n\to\infty} \hat{\Theta}, \quad Z_n = (M_n \circ \hat{D}_n), \quad \tilde{Z}_n = X_n(X_n'X_n)^{-1}X_n'Z,$$
$$\Theta_2 = \lim_{n\to\infty} \frac{1}{n}\Big(I - Z_n\hat{\Theta}\tilde{Z}_n'/n\Big)' X_n(X_n'X_n)^{-1}X_n'\Big(I - Z_n\hat{\Theta}\tilde{Z}_n'/n\Big)$$

Theorem 3 shows that the 2SLSS estimator achieves normality at the stan-

dard rate $\sqrt{n}$. The shifts Δ_1 and Δ_2 represent the bias from using nodewise regression and they are shown to be $o_p(1)$ with the proper choice of tuning parameters.

Corollary 2. *In the heterogeneous endogenous effects model with cliques and under assumption 1′-3′, and assumptions 4-5, if* $\lambda \propto \sqrt{\log n/n}$

$$\sqrt{n}\begin{pmatrix}(\hat{e}-\eta_0)\\(\hat{r}-\gamma_0)\end{pmatrix} = E_1 + \Delta_1,$$

$$\sqrt{n}(\hat{b}-\beta_0) = E_2 + \Delta_2,$$

where

$$E_1 \sim N(0, \sigma^2\Theta_1 diag(\Gamma)\Omega diag(\Gamma)\Theta_1'),$$

$$E_2 \sim N(0, \sigma^2\Theta_2 diag(\Gamma)\Omega diag(\Gamma)\Theta_2'),$$

and

$$\|\Delta_1\|_\infty = o_p(1), \quad \|\Delta_2\|_\infty = o_p(1),$$

$$\Gamma = \lim_{n\to\infty}(I - M_n \circ \eta_0)^- X_n\beta_0,$$

$$\Theta_1 = \lim_{n\to\infty}\hat{\Theta}, \quad Z_n = [(M_n \circ \hat{D}_n), M_n\hat{D}_n], \quad \tilde{Z}_n = X_n(X_n'X_n)^{-1}X_n'Z,$$

$$\Theta_2 = \lim_{n\to\infty}\frac{1}{n}\Big(I - Z_n\hat{\Theta}\tilde{Z}_n'/n\Big)' X_n(X_n'X_n)^{-1}X_n'\Big(I - Z_n\hat{\Theta}\tilde{Z}_n'/n\Big)$$

For my setting with multiple networks, I derive the following results:

Theorem 4. *In the heterogeneous endogenous effects model with multiple networks*

and under assumptions 1-5*, if* $\lambda_1 \propto \sqrt{\frac{\log n}{n}}$ *and* $\lambda_2 \propto \sqrt{\frac{\log n}{n}}$

$$\sqrt{n}(\hat{e}_m - \eta_0) = E_{m1} + \Delta_{m1},$$

$$\sqrt{n}(\hat{b}_m - \beta_0) = E_{m2} + \Delta_{m2},$$

where

$$E_{m1} \sim N(0, \sigma^2 \Theta_{Z1} diag(\Gamma) \Omega_m diag(\Gamma) \Theta'_{Z2}),$$

$$E_{m2} \sim N(0, \sigma^2 \Theta_{Z2} diag(\Gamma) \Omega_m diag(\Gamma) \Theta'_{Z2}),$$

and

$$\|\Delta_{m1}\|_\infty = o_p(1), \quad \|\Delta_{m2}\|_\infty = o_p(1),$$

$$\Theta_{Z1} = \lim_{n\to\infty} \hat{\Theta}_Z, \quad Z_n = (M_n \circ \hat{D}_n), \quad \tilde{Z}_n = X_n (X_n' X_n)^{-1} X_n' Z,$$

$$\Theta_{Z2} = \lim_{n\to\infty} \frac{1}{n} \Big(I - Z_n \hat{\Theta} \tilde{Z}_n'/n\Big)' X_n (X_n' X_n)^{-1} X_n' \Big(I - Z_n \hat{\Theta} \tilde{Z}_n'/n\Big)$$

The proof of Theorem 2 and Theorem 4 requires the following results from the square-root sparse group LASSO: 1) Bounds on the prediction, i.e. $\left\| \sum_{j=1}^q (M^j \circ X_n)(\hat{\eta}^j - \eta_0^j) + X_n(\hat{\beta} - \beta_0) \right\|_2 \lesssim \lambda.$ and 2) Consistency of selection i.e. $\hat{S}_n = S$. I prove these two statistical properties in the appendix.

1.6 Simulations

In this section, I report Monte Carlo simulation results for my heterogeneous endogenous effects model and its extension with cliques and with multiple networks. My results are robust when applied to networks generated by different algorithms and to networks of different sizes.

1.6.1 Heterogeneous Endogenous Effects Model

To assess the finite sample performance of my estimator for the heterogeneous endogenous effects model, I use the Erdos-Renyi algorithm to simulate a network of size n. Individuals are added into the graph one at a time. When one individual is added to the network, she has probability p of generating a link with all existing individuals independently. I choose $p = 0.1$ and $p = 0.2$ in the simulation. I avoid a large p because collinearity among regressors may arise when links become very dense, violating assumption 5.

I set the first 5 individuals to be influential by letting their coefficients η_j be non-zero. To guarantee the existence of endogenous effects, I arbitrarily specify the connections among these five individuals. The adjacency matrix M_n for the five influential individuals is given in the appendix. If the connections among these five individuals are not fixed, there is a possibility that no connections are formed among these five and thus there is no endogeneity in the network. In this case, the results will be too good in such a case.

The true parameters are fixed as $\beta_0 = 3$, $\eta_{0,1} = \eta_{0,2} = \eta_{0,3} = \eta_{0,4} = \eta_{0,5} = 0.5$, and $\eta_{0,j} = 0$ for $j > 5$. Individual characteristics X_n are generated from a standard normal distribution.

Individual outcomes Y_n are then generated as $Y_n = (I - M_n \circ \eta_0)^{-1}(X_n\beta_0 + \epsilon_n)$ where ϵ_n is drawn independently from a standard normal distribution.

I use (M_n, X_n, Y_n) as observations and apply my two-stage LASSO estimator. I construct the de-sparse 2SLSS estimator and repeat the above process 200 times in a manner similar to [89].

I report the average coverage probability (Avgcov) and average length (Avglength) of confidence intervals for the coefficients for influential individuals, $\{\eta_1, \cdots, \eta_5\}$, the coefficient for individual characteristics, β_0, and the coefficients for non-influential individuals, the η_js ($j > 5$). For example:

$$\text{Avgcov } S_0 = s_0^{-1} \sum_{j \in S_0} \mathbb{P}[\eta_{0,j} \in CI_j] \tag{1.14}$$

$$\text{Avglength } S_0 = s_0^{-1} \sum_{j \in S_0} length(CI_j) \tag{1.15}$$

I separately report the average coverage and average length for each of the five influential individuals. As shown in table A.1, the coverage is around the nominal 95% level and the length of the confidence intervals decreases as the sample size grows.

Since we can construct confidence intervals for all n coefficients, joint inference can be performed under the control of False Discover Rate (FDR). As shown in equation (1.16), the power reported in table A.1 represents the average percentage in the active set (i.e. $\{1, 2, 3, 4, 5\}$) that is significant after controlling for the False Discover Rate (FDR) at 5% using the Benjamini-Hochberg method. The FDR reported in table A.1 represents the average percentage of the non-active set (i.e. $\{6, 7, \cdots, n\}$) that is significant after controlling the FDR at 5% using the Benjamini-Hochberg method. The exact definition is as in equation (1.17).

$$\text{Power } = s_0^{-1} \sum_{j \in S_0} \mathbb{P}[H_{0,j} \text{ is rejected}] \tag{1.16}$$

$$\text{FDR } = \sum_{j \in S_0^c} \mathbb{P}[H_{0,j} \text{ is rejected}] / \sum_{j=1}^{n} \mathbb{P}[H_{0,j} \text{ is rejected}] \tag{1.17}$$

The power varies because the networks change when the sample size increases. It is strictly increasing when the network is sparse (i.e. $p = 0.1$). The power decreases in the $p = 0.2$ case as the problem of endogeneity increases when the network is dense. The empirical FDR is controlled well, which all under the 5% rate. Notice that the confidence interval's length is large when the sample size equals 50. This is because when the number of individuals is small, some individuals might only connect to 1 or 2 other individuals. This means that the regressors that represent this individual are all 0s except for a small numbers of non-zero terms, which leads to a large standard error.

The two-stage LASSO estimator requires the choice of two tuning parameters (i.e. the two λs from both stages as in section 4.1). Moreover, when calculating $\hat{\Theta}$ in the De-sparse 2SLSS estimator (section 4.2) and using the nodewise regression, one also need to choose a tuning parameter. Following the suggestion in [13], I use a benchmark choice of λ for the first stage and nodewise regression (i.e. $\lambda \propto \Phi^{-1}(1 - \alpha/(2n))/\sqrt{n}$), where $\Phi^{-1}(.)$ is the inverse of normal cdf function. For the second stage, I use cross-validation to pick λ to enhance finite sample performance.

I further increase the number of influential individuals to 10 and report the results in table A.2. Again, to guarantee the existence of endogeneity, the adjacency matrix for these ten individuals is set as shown in the appendix. All average coverages and average confidence interval lengths are separately reported for these ten individuals.

The choice of the tuning parameters is similar to those used to generate table A.1 for networks with 50 and 200 individuals. For networks with 500 individuals, I use benchmark λ to replace cross validation in the second stage. The

idea is to show the converge of the process, such that valid coverage can still be generated under theory guide tuning parameters [see 13].

As shown in table A.2, all coverages are very close to the nominal levels. The average lengths of confidence intervals is slightly larger compared with table A.1. This is due to the increase in influential individuals; it is more difficult to differentiate them from those irrelevant individuals.

Table A.3 presents the result when a network is generated using the Watts-Strogatz mechanism or the "small world" network. Define the pN (even number) as the mean degree for each node and a special parameter $\omega = 0.4$. The WattsStrogatz mechanism works as follows:

- construct a graph with N nodes each connected to pN neighbors, which $\frac{pN}{2}$ on each side.
- For each node n_i, take every edge (n_i, n_j) with $i < j$ and rewrite it with probability ω. Rewrite means replace (n_i, n_j) with (n_i, n_k) where k is choosing uniformly among all nodes that are not currently connected with n_i

The influential individuals are chosen as the 1st, 5th, 15th, 40th and 50th individuals in the network. As shown in table A.3, my estimator is robust under a "small world" algorithm. Nominal level is reached as the size of the network grows and the length of confidence intervals is slightly smaller than in the standard case.

1.6.2 Heterogeneous Endogenous Effects Model with Cliques

Table A.5 presents results for the heterogeneous endogenous effects model with cliques. The outcome variable Y_n is now generated as $Y_n = (I - M_n \circ \eta_0 - M_n\gamma_0)^{-1}(X_n\beta_0 + \epsilon_n)$. The coefficient of the homogeneity effects γ_0 is set at 0.05.

The choice of the tuning parameters is similar to that used to generate table A.1 for networks with 50 and 200 individuals. For networks with 500 individuals, I use benchmark λ (i.e. $\lambda \propto \Phi^{-1}(1 - \alpha/(2n))/\sqrt{n}$) to replace cross validation in the second stage.

The coverage is above the 95% nominal level in all cases. I also report the mean coverage and average length of the confidence interval for the coefficient of the homogeneous effects. My model gives above 95% coverage in all cases. I also report the empirical probability of rejecting a null hypothesis of zeros effects at 95% nominal level. The probability of rejecting the test converges to 1 when the sample size grows to 500.

1.6.3 Heterogeneous Endogenous Effects Model with Multiple Networks

In this Monte Carlo exercise, I include two different networks generated by the Erdos-Renyi algorithm, where one is influential and the other is not. I use the two-stage LASSO estimator with multiple networks to estimate the parameters. The square-root sparse group LASSO requires two tuning parameters, one for the l_2 norm and the other for the l_1 norm. I set the two parameters to be equal to each other as the correlations among the columns of the adjacency matrices

are very small. The choice of tuning parameters is similar to that used to generate table 1 for networks with 50 and 200 individuals. For networks with 500 individuals, I use a rule of thumb to choose λ instead of cross-validation in the second stage. Table A.4 summarizes the results. As in previous results, all coverages exceed the nominal 95% level.

I report the empirical probabilities such that at least one individual is detected in a given network controlling for the FDR at 5% using the Benjamini-Hochberg method. I also report the average number of detections conditioning on at least one individual who is detected in a given network. Tables A.4 shows that network 1, which is the relevant network, is more likely to be detected in all cases than network 2, the irrelevant network. The average number of identified individuals for network 1 is also more than that of network 2.

1.7 Empirical Application

I use the proposed estimator to study the importance of different networks in spreading the participation in a micro finance program within rural Indian villages. I show that different kinds of networks have different effects on individuals decisions. I identify the influential individuals in each village. My analysis shows that leaders among agricultural laborers, Anganavadi teachers, construction workers, small business owners and mechanics are very likely to be influential in the villages.

1.7.1 Background

A non-profit organization named Bharatha Swamukti Samsthe (BSS) has been running micro finance programs in rural southern Karnataka, India since 2007. It provides small loan products to poor women and, through them, to their families. The villages covered by the program are geographically isolated and heterogeneous in terms of caste.

When BSS initially introduces a micro finance program to a village, the credit officers of BSS first approached a number of "predefined leaders", such as teachers, shopkeepers and village elders. BSS held a private meeting with these leaders and explained the program. Then these predefined leaders passed the information onto other villagers. Those who were interested in the program and contacted BSS were trained and assigned to groups to receive credit. Each group consisted of 5 borrowers and group members were jointly liable for loans. Loans were around 10,000 rupees (approximately $200) at an annualized rate of approximately 28%. Note that 74.5 percent of the households in rural area said the monthly income of their highest earning member is less than 5,000 rupees (source: Socio-Economic Caste Census-2011). This loan had to be repaid within 50 weeks.

In 2006, 75 villages in Karnataka were surveyed 6 months before the initiation of the BSS micro finance program. This survey consisted of a village questionnaire and a detailed follow-up survey conducted among a subsample of villagers. The village questionnaire gathered demographic information on all households in a village including GPS coordinates, age, gender, number of rooms, whether the house had electricity, and whether the house had a latrine. The data set also contains information on the "pre-defined leaders" set who

helped spread the information to the entire village. The follow-up survey collected data from a villager sample stratified according to age, education level, caste, occupancy, etc. It also asked questions about social network structures along 12 dimensions, including:

- Friends: Name the 4 non-relatives whom you speak to the most.
- Visit-go: In your free time, whose house do you visit?
- Visit-come: Who visits your house in his or her free time?
- Borrow-kerorice: If you needed to borrow kerosene or rice, to whom would you go?
- Lend-kerorice: Who would come to you if he/she needed to borrow kerosene or rice?
- Borrow-money: If you suddenly needed to borrow Rs. 50 for a day, whom would you ask?
- Lend-money: Who do you trust enough that if he/she needed to borrow Rs. 50 for a day you would lend it to him/her?
- Advice-come: Who comes to you for advice?
- Advice-go: If you had to make a difficult personal decision, whom would you ask for advice?
- Medical-help: If you had a medical emergency and were alone at home whom would you ask for help in getting to a hospital?
- Relatives: Name any close relatives, aside from those in this household, who also live in this village.
- Temple-company: Do you visit a temple/mosque/church? Do you go with anyone else? What are the names of these people?

For the 43 villages where micro finance was introduced by the time of 2011, BSS also collects information on which villagers have joined the program. These survey questions reveal the underlying structures for connections among any two individuals in the network. Figure 1.4 presents all those connections at the household-level in a graph. Each node in the graph represents a household. A green node indicates that the household joined the micro finance program, while a blue node indicates that it did not. Bigger nodes represent those households in which at least one family member has been chosen as being among the "pre-defined leaders". An edge between two nodes signifies that the two nodes are connected in at least one of the 12 networks. The darker the color of the edge, the more connections it represents.

This dataset provides an ideal framework for application of the heterogeneous endogenous effects model. First, it allows me to model endogenous effects. An individual may decide to join the micro finance program if her neighbors or friends plan to join. Second, the endogenous effects are individual specific. Given the diversity of the villagers, it is possible that some villagers are more influential than others. Third, it allows me to implement the heterogeneous endogenous effects model with multiple networks. The questions asked regarding multiple dimensions of the network structure allow me to explore which network is most influential.

1.7.2 Data

In this empirical study, I focus on the 38 villages that have been introduced to the micro finance programs by BSS and have data publicly available [3]. For each

[3]The dataset can be downloaded from http://web.stanford.edu/ jacksonm/Data.html

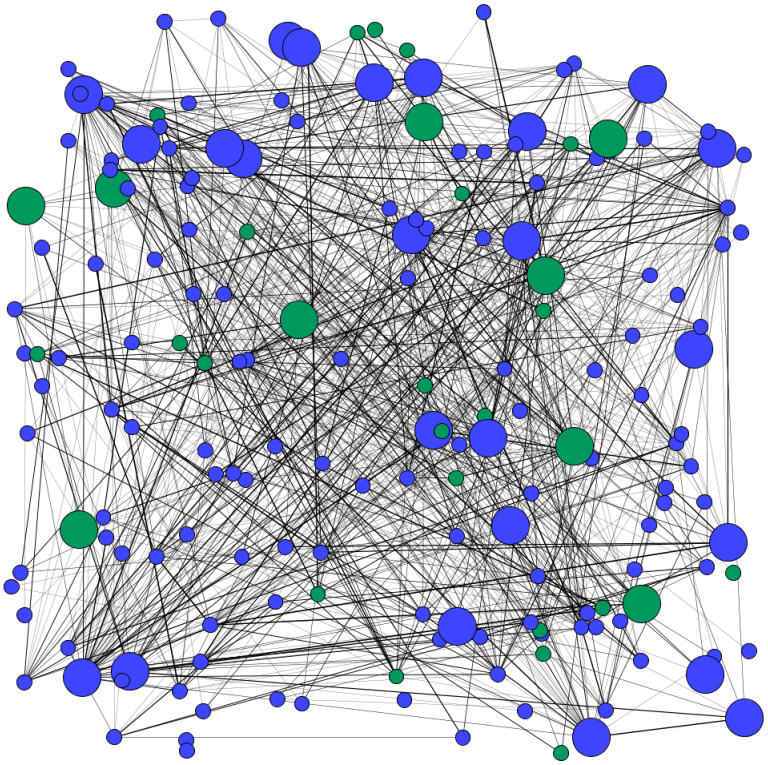


Figure 1.4: Network in Village 1

village, I can observe both its social network structure and the villagers' decisions about joining the program. I drop the data for one village (Village 46) that contains incorrect entries on the index of households. Table A.6 summarizes the descriptive statistics for each village.

Among the 12 questions about the social network structure, 4 pairs essentially capture the same connections among the villagers [4]. Therefore, I consolidate each pair of questions into one dimension:

[4]Assuming every villager truthfully answers a pair of questions, the adjacency matrices associated with each question are the same. It is also plausible to treat villagers' answers to each question as a separate directed graph. However, these questions do not allow for clear determination of the directions. For example, if villager A visits villager B's house, it is not clear whether villager A influences villager B or vice versa

- Visit-go-come
  - In your free time, whose house do you visit?
  - Who visits your house in his or her free time?
- Borrow-Lend-kerorice
  - If you needed to borrow kerosene or rice, to whom would you go?
  - Who would come to you if he/she needed to borrow kerosene or rice?
- Borrow-Lend-money
  - If you suddenly needed to borrow Rs. 50 for a day, whom would you ask?
  - Who do you trust enough that if he/she needed to borrow Rs. 50 for a day you would lend it to him/her?
- Help decision
  - Who comes to you for advice?
  - If you had to make a difficult personal decision, whom would you ask for advice?

I restructure all the data at the household level as only women are allowed to apply for the micro finance program because the goal of BSS is to support families through the women in them. As a result, a womans decision to join or not join the micro finance program becomes her familys decision. A connection between two villagers becomes a connection between two families. A "predefined leader" is a villager selected by BSS to help spread information about the micro finance program to the other villagers. At the household level, I use the term "predefined leader" for a household that contains at least one such villager.

1.7.3 Sparsity and Equilibrium

To demonstrate how my method identifies influential households, I model families' decisions regarding joining the micro finance program as a network game with Bayesian Nash Equilibrium. For household i, let d_i^* be the expected probability that i chooses to join the micro finance program. The decision of household i depends on its neighbors' decisions as well as the types of connections between them. The decision also depends on its characteristics X_i and on unobserved information ϵ_i. Formally, it can be written as:

$$d_i^* = \sum_{l \in N_i} d_l^* (\sum_{j=1}^{q} \eta_l^j) + x_i\beta + \epsilon_i$$

Rewritten in matrix form:

$$D_n^* = \sum_{j=1}^{q} \left(M_n^j \circ D_n^* \right) \eta^j + X_n\beta + \epsilon_n$$

By Assumption 1-3, there is a unique equilibrium that determines D_n.

I assume that only a small number of households are influential over their neighbors. Leaders and followers are usually observed in those rural villages. Big decisions are often made by the village elders or by the more educated among the villagers. BSS recognized the importance of leaders and gathered a group of predefined leaders, asking them to inform the rest of the villagers about their program. I do not consider the local level influence in these villages given the size and how complicated the network structures are. Households are closely connected by these 8 networks as shown in Figure 1.4 and there is no form of clique visible.

Because the villages are considered geographically isolated, I apply my estimator separately to each of the 38 villages. I use the number of rooms per person

in a houshold as the independent variable X_n. Number of rooms per person is a proxy for the wealth in the family. As shown in table 1.1, it is negatively correlated with the decision to join the micro finance program. The richer the family, the less likely the family is to participate in the micro-finance program. I further check the robustness of my independent variable by including additional controls. The adjacency matrix M_n^j is constructed from the questions in the survey. Households i and k are connected in network j if either i or k reported the other in question j. Finally, d_i^* is replaced with the household's choice.

The instruments are constructed as $\left(M_n^j \circ X_n\right)$ for $j = 1, 2, \cdots, 8$. I use the heterogeneous endogenous effects model with multiple networks to: 1) Identify the effective networks affecting a household's decision and 2) Identify that households that are leaders in the village and study the association between observable characteristics and leader status. If a new program is going to try to recruit these households, the organizers can target those influential households and try to persuade them to join first.

1.7.4 Results

Identifying Effective Networks

First, I study how LASSO selects networks. I define a coefficient for a household's endogenous effect in a network as significant according to two different criteria. The first criterion, "Cross-Validation", determines a coefficient to be significant if LASSO predicts the coefficient to be non-zero after cross-validation. The second criterion, "De-sparse", first constructs a bias-adjusted coefficient and calculates its standard error. It then determines a coefficient to be signifi-

Table 1.1: Predictive Power of Characteristics X_n

	(1) Decision	(2) Decision	(3) Decision
Average # room	-0.0832***	-0.0717***	-0.0343***
	(0.0091)	(0.0101)	(0.0109)
Household Size		0.0048**	0.0041**
		(0.0019)	(0.0020)
Electricity			0.0166
			(0.0150)
Latrine			-0.0539***
			(0.0093)
Average #workers			0.0064*
			(0.0039)
Average age			-0.0028***
			(0.0004)
n	8,375	8,375	8,375
R^2	0.0477	0.0484	0.0591

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable is households' decision on whether to join the micro finance program or not. All design control village fixed effects.

cant if the Benjamini-Hochberg method rejects the null hypothesis of zero effect at the 5% false discovery rate. A network is defined as significant if at least one coefficient for heterogeneous endogenous effects in this network is significant.

Table 1.2 presents the empirical probability of the 8 networks being significant among the 37 villages. Note that certain types of networks (such as visit go-come) are more likely to pass influence then others (such as temple company). For example, by cross-validation criterion, the visit go-come network is detected as significant in 19 out of the 37 villages (i.e. 51%) whereas temple company is detected as significant in only 5 out of the 37 villages (i.e. 14%). I also present the average number of households associated with significant endogenous effects in each significant network. For example, according to the cross-validation criterion, 342 households in 19 villages have significant coefficients associated with the visit go-come network, which averages to 18 households per detection. On the other hand, 32 households in 5 villages have significant coefficients associated with the temple company network, which averages to 6 households per detection.

In terms of variable selection, if Assumption 4 holds, the cross-validation criterion may consistently select the truly influential households with high probability even in a finite sample. On the other hand, the de-sparse criterion is likely to be conservative because of its use of the false discovery control process. In terms of coefficients estimated, de-sparse estimators are asymptotically consistent. On the other hand, estimates based on the LASSO estimator suffer from shrinkage bias and are not consistent.

Table 1.3 reports the average absolute heterogeneous endogenous effects within significant networks using the de-sparse estimators. For example, if all

else is equal, an additional influential neighbor in the visit go-come network will, on average, increase the probability of joining the micro-finance program by 16%; moreover, an additional influential neighbor in both the visit go-come network and the friendship network will increase the probability of joining the micro-finance program by 16% on average. The magnitude of those coefficients should not be over interpreted as exogenous effects and correlated effects are not considered in the model. Similar to Table 1.2, certain types of networks (such as visit go-come) pass stronger influence than others (such as temple company). Note that, in most of the cases, networks that are more likely to pass influence also pass stronger influence. The relative network is an exception. Even though the relative network is less likely to pass influence compared to the friendship network, the borrow-lend-money network and the help decision network, it passes stronger influence once it is significant. Table 1.3 also presents the percentage of positive effects detected among different networks. For networks such as visit-go-come and friendship, more than 70% of influential villagers are "true leaders" – if they decide to join the micro-finance program, their neighbors will follow them and join the program. On the contrary, for the temple company network, it is almost equally likely for neighbors of influential households to either follow the same decision or choose the opposite.

[insert table 1.2]

[insert table 1.3]

The results in Table 1.2 and Table 1.3 suggest villagers are more likely to discuss the micro-finance program when they visit each other, chat with friends,

and meet with people to whom they are economically connected. Villagers are not likely to talk about the micro finance program when they go to the temple.

To verify my findings above, I provide exogenous evidence using centrality measures. Intuitively, the more a villager is exposed to a network, the more likely she is to be connected to influential villagers, and hence she is more likely to join the program. Following [9], I measure the centrality of each villager in each network through "degree", "closeness", "betweenness" and "eigenvector". (See Appendix for definitions) Then I regress households' decisions on whether to join the micro finance program on each centrality measure separately while controlling for village fixed effects:

$$d_j = C_j^q \beta + \gamma_j + \epsilon_j \tag{1.18}$$

where d_j is household j's decision; C_j^q is household j's centrality in the network q; γ_j is the village fixed effect; and ϵ_j is the error term.

Table 1.7 presents the regression results for equation (1.18). Visit go-come and borrow-lend kerorice are positively correlated with degree, closeness and eigenvector centrality. Friendship, borrow-lend money and medical help are positively correlated with degree and closeness centrality. This is consistent with my findings that these four networks are more effective in passing influence. Meanwhile, neither help decision, relative nor temple company are found to be correlated with any of the centrality measures defined above. This is also consistent with the lower probability of passing influence as found in table 1.2. Note that none of the networks are found to correlate with betweenness centrality. This is because betweenness centrality is based on the shortest paths in a network, which is not a direct measure of the exposure of an individual to a network.

[insert table 1.6]

Identifying Influential Households

Second, I focus on how LASSO selects households. I compare the LASSO selected influential households with the BSS selected "predefined leaders". It is important to point out that these "predefined leaders" are *not* necessarily influential villagers in a network. Recall that predefined leaders are a set of villagers that BSS select to help spread the information about the micro finance program. The fact that a villager is selected as a "predefined leader" to *pass information* about the micro finance program does not a priori guarantee her or her family's *influence* – her decision to join the micro finance program may not lead to her neighbors' decisions to join. In the analyses below, I will examine how influential villagers are associated with "predefined leaders" and explore their potential differences.

1. Influential Predefined Households

In table 1.7, I report results indicating that influential households selected by LASSO partly overlap with "predefined leaders". This is intuitive because some "predefined leaders" such as school headmasters and village elders are highly respected figures in a village. Therefore, their decisions are likely to be followed by others in the village. On average, BSS selected 27 villagers as "predefined leaders" in each village. In comparison, Cross-Validation criterion selects around 22 villagers and De-sparse criterion selects around 6. Furthermore, on average, 4 out of 22 influential villagers (i.e. 19%) selected by Cross-Validation criterion are also BSS "predefined leaders"; 1 out of 6 influential villagers (i.e. 13%) selected by De-sparse criterion are also BSS "predefined leaders". In Table

A.7 below, I show that small business owners are more likely to be both influential and selected as "predefined leaders".

[insert table 1.7]

2. Influential Non-Predefined Households

In this and the following section, I focus on understanding the differences between the influential households selected by LASSO and the "predefined households" selected by BSS. I investigate the likelihood that a household being selected by LASSO or by BSS, as associated with the careers of its family members. More specifically, I regress whether a household is selected as "predefined leader" (Design (1)), whether a household is selected by LASSO as influential (Design (2)), and whether a household joins the micro finance program (Design (3)), separately on dummy variables based on the full set of careers as reported in the survey data controlling for other household characteristics and village fixed effects. The full results of these regressions are reported in Table A.7 in the Appendix.

Table 1.6 summarizes all careers that have a significant impact on the likelihood of a household being selected by LASSO as influential. Note that except for small business owners, all the other careers in this table are not significantly associated with the likelihood of a household being selected by BSS as being among the "predefined leaders". Over 67% of the villagers are agricultural laborers and 75% of the LASSO selected influential households have agricultural laborers in the family. Anganwadi Teacher is a set of groups that provides preschool education to the children. They are part of the government's health care system in the rural areas. There are 31 Anganwadi Teachers in all villages, and

LASSO detects 7 of their families to be influential. BSS also selects 7 of them as "predefined leaders" but only 2 of the 7 are selected by LASSO as influential. Other careers that are correlated with LASSO selection include police officer, mechanic, and skilled laborers. These are more educated individuals and it seems compelling that they are selected as influential individuals.

Table 1.7 summarizes all careers that have a significant impact on the likelihood of a household being selected by BSS as being among the "predefined leaders". Poojari are Indian priests in those villages and they are very likely to be included as "predefined leaders". However, they are not likely to influence people to join the micro finance program. Other careers as tailor, hotel workers, veteran, and barber are included as "predefined leaders" because individuals doing these jobs can spread information quickly in the village. However, LASSO does not find these individuals to be influence.

[insert table 9]

1.8 Conclusions

In this paper, I propose a novel SAR model which allows for *heterogeneous* endogenous effects. Specifically, each individual has an individual-specific endogenous effect on her neighbors. My approach is useful for modeling a network with leaders and followers. For example, it can model how online opinion leaders influence the public or how experienced workers boost coworkers' productivity.

I propose a set of instruments as well as a two stage LASSO (2SLSS) method to estimate my model. The instruments are constructed as a function of the independent variables and an adjacency matrix. I use a LASSO type estimator to select the valid instruments in the first stage and the influential individuals in the second stage. I propose a bias correction for my two-stage estimator following [89]. I derive the asymptotic normality for my "de-sparse" two-stage LASSO estimator and conduct robust inference including confidence intervals.

My model can be extended to allow for more flexible structures. To apply LASSO, I assume that the number of influential individuals is sparse. I propose heterogeneous endogenous effects model with cliques to incorporate locally influential individuals, where the sparsity assumption is only applied to globally influential individuals. My model can also be extended to situations where there are multiple networks. I propose the use of the square-root sparse group LASSO in my 2SLSS process. I derive the convergence rate and prove the consistency of selection for the square-root sparse group LASSO estimator.

I apply my method to study villagers' decisions to participate in microfinance programs in rural areas of Indian. I show that leaders in those villages have significant influence over their neighbors' decision to join the microfinance program, and I provide rankings for the different social and economic networks among villagers. Based on how effectively each network spreads the impact of influential individuals' decisions, my method shows that some networks such as "visit go-come" and "borrow money" are much more effective in influencing villagers' decisions than other networks such as "temple company" and "medical help". I further show that individuals from certain careers such as agricultural workers, Anganwadi teachers and small business owners are more

likely to influence other villagers.

There are two interesting directions for future research. First, it is possible to include heterogeneous exogenous effects in the model. These effects aim to capture how an individual's outcome varies with the exogenous characteristics of her neighbors. However, when both exogenous and endogenous effects are included in standard SARs, an identification problem known as the "reflection problem" may arise [see 74]. A similar problem arises also in my model if heterogeneous exogenous effects are included. [19] show that under additional assumptions on the adjacency matrix, this problem can be solved in SARs. With similar restrictions on the adjacency matrix, it is possible to construct a new set of instruments to include heterogeneous exogenous effects in my model.

Second, it might be possible to use penalized GMM type estimator to estimate my model. 2SLS and GMM are the two most commonly used estimators to deal with endogeneity in SARs. My 2SLSS can be rewritten as a penalized GMM problem. The current progress on penalized GMM estimators include [41] and [70]. But no uniformly valid inference method currently exists for penalized GMM.

Table 1.2: Second Stage: network usage

		visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
Cross [1]	probability [3]	51%	43%	41%	41%	30%	32%	30%	14%
Validation	identified [4]	18	12	13	14	9	14	9	6
De-sparse [2]	probability [3]	51%	46%	51%	43%	32%	41%	43%	19%
	identified [4]	3	3	3	2	3	3	3	2

0. Reported are the probability of detection among the 38 villages.

1. Cross Validation represents those networks identified from lasso using cross validation.

2. De-sparse represents those networks identified from De-sparse criterion using FDR control.

3. Probability reports the empirical probability that at least one regressor in the group is significant.

4. Identified reports the averaged number of significant regressors in the group conditioning on the network being significant.

Table 1.3: Second Stage: average endogenous effect

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
absolute magnitudes [1]	0.1543	0.1443	0.1245	0.1194	0.1214	0.1217	0.1404	0.0555
percentage of positive effect	77%	67%	69%	70%	68%	77%	67%	55%

1. The magnitudes are reported based on the de-sparse estimator.

Reported numbers are conditioning being significant using De-sparse method.

Table 1.4: Centrality Measure

	visit go-come	borrow-lend kerorice	borrow-lend money	friendship	medical help	help decision	relatives	temple company
degree	0.0025**	0.0032**	0.0020**	0.0022**	0.0032**	0.0013	0.0035	0.0061
	(0.0009)	(0.0011)	(0.0010)	(0.0010)	(0.0014)	(0.0011)	(0.0019)	(0.0032)
closeness	32.9116***	40.2695***	29.7981**	31.0882***	32.5602**	18.1944	7.9509	231.0770
	(9.5639)	(10.7603)	(9.3901)	(9.0446)	(11.1383)	(9.8242)	(16.5557)	(134.3147)
betweenness	1.3565	0.1751	0.2940	1.6713	1.1662	-0.5728	0.3055	-0.2093
	(1.0101)	(0.8240)	(0.9504)	(1.0207)	(0.8634)	(0.8283)	(0.7736)	(0.2178)
eigenvector	3.6201***	1.5239**	0.1161	1.3927	-0.7338	-0.2387	0.7753	3.3015
	(0.8888)	(0.6250)	(0.8245)	(0.8271)	(0.7709)	(0.7603)	(0.5672)	(3.5585)

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Definition for centrality measures are in appendix.

Table 1.5: Second Stage: coverage of predefined leaders

	coverage [2]	total number of discovery [3]
Cross Validation [4]	19%	22
De-sparse [5]	13%	6

1. predefined leaders are a set of villagers defined by BSS, who helped spread the information about the micro-finance program.

2. Coverage reports the percentage of individuals detected by LASSO and also selected as "predefined leaders" in total detection.

3. Total number of discovery reports the total number of individuals discovered by lasso using each method.

4. Cross Validation represents those individuals identified from lasso using cross validation .

5. De-sparse represents those individuals identified from De-sparse criterion controlling FDR.

6. The average number of predefined leaders in one village is 27

Table 1.6: Second Stage: who are they

	(1)	(2)	(3)
Agriculture labour	-0.0141	0.0476*	0.0672***
	(0.0136)	(0.0286)	(0.0134)
Anganwadi Teacher	0.0386	0.0664	0.1248**
	(0.0602)	(0.1269)	(0.0593)
Blacksmith	-0.0752	-0.2279	0.1606*
	(0.0927)	(0.1954)	(0.0913)
Construction/mud work	0.0050	0.2199***	0.0562**
	(0.0258)	(0.0544)	(0.0254)
Small business	0.2006***	0.1287***	0.0606***
	(0.0227)	(0.0479)	(0.0224)
Police officer	-0.1459	-0.0374	0.3282*
	(0.1917)	(0.4044)	(0.1890)
Mechanic	0.0106	-0.1237	0.1274**
	(0.0634)	(0.1337)	(0.0625)
Skilled labour/work for company	0.0469	0.0252	0.0809*
	(0.0491)	(0.1036)	(0.0484)
Control other careers	Y	Y	Y
Control village fix effect	Y	Y	Y

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

design (1) uses whether one is predefined leaders as response variable

design (2) uses whether one joins the micro-finance program as response variable

design (3) uses whether one is selected by lasso as response variable

Table 1.7: Second Stage: who are they

	(1)	(2)	(3)
Small business	0.2006***	0.1287***	0.0606***
	(0.0227)	(0.0479)	(0.0224)
Tailor Garment worker	0.0903***	0.1169*	0.0309
	(0.0304)	(0.0642)	(0.0300)
Hotel worker	0.3299***	0.4257***	0.0759
	(0.0750)	(0.1581)	(0.0739)
Poojari	0.3697***	-0.1542	0.1501
	(0.1369)	(0.2887)	(0.1349)
Veterinary clinic	0.8649***	1.9114***	0.0377
	(0.3314)	(0.6990)	(0.3266)
Barber/saloon	0.4883***	-0.0036	0.0443
	(0.1005)	(0.2119)	(0.0990)
Doctor/Health assistant	0.2691**	0.2703	0.0874
	(0.1053)	(0.2222)	(0.1038)
Control other careers	Y	Y	Y
Control village fix effect	Y	Y	Y

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

design (1) uses whether one is predefined leaders as response variable

design (2) uses whether one joins the micro-finance program as response variable

design (3) uses whether one is selected by lasso as response variable

CHAPTER 2

ON TESTING CONTINUITY AND THE DETECTION OF FAILURES

2.1 Introduction

This paper introduces a method for detecting discontinuities for when the econometrician does not know the underlying parametric form, the number of discontinuities, the location of discontinuities, or their type (point, jump, kink, etc). Often detection of these discontinuities is of direct economic interest: e.g., identification of tipping points in demographic changes [24] or cheap-talk signaling conventions [6]. In the most general sense, however, it is a specification test: both a test of continuity as well as an opportunity to learn about discontinuities from the data in an agnostic, nonparametric way.

Our agnosticism implies that we are testing against a large set of alternative hypotheses – many possible breaks – and so our procedure is adaptive. We are openly engaging in "data snooping," i.e. model selection, and it is well-known that this invalidates standard inference. Inference after model selection, respecting these limitations, is an area of great recent interest. Our solution is to build the desired inferential guarantees into the model selection procedure itself. In particular, we develop an estimator that detects discontinuities while controlling the False Discovery Rate (FDR) – i.e. the ratio of type I errors to the total number of rejections – at a pre-specified level [14].

The model selection component of our algorithm uses a LASSO framework to construct an ordered sequence of hypothesis tests, i.e. potential discontinuities. We construct conditionally valid test statistics for each – i.e. we explicitly

condition on the event the hypothesis test was selected. [42] demonstrates that this conditioning is closely related to – and more powerful than – data splitting. Because each hypothesis test is conditional on the last, we must then solve a sequential multiple comparison problem (MCP). To this end we take advantage of the recently developed Forward Stop algorithm of [46] which extends FDR control from the simultaneous MCP setting of [14] to the sequential MCP setting. In simulations we are able to show that this sequential approach is more powerful than using standard FDR control with confidence intervals proposed by other recent work [89, 12]. Intuitively, this is because that approach ignores correlation between the test statistics, which ours exploits. We somewhat incidentally contribute, therefore, to an emerging literature on using that correlation in FDR control.

We also draw on a large literature on the detection of structural breaks. There are a number of salient technical challenges that his literature has confronted, beginning with the multiple comparison problem (MCP) implicit in testing at many potential break locations. The solution, to formalize the MCP as an order statistic problem, was proposed by [51] and generalized by [4]. A second thread in this literature has focused on the construction of confidence intervals for the parameter governing the size of a break at the most likely location [48, 50]. This becomes a nonstandard inference problem because nesting the null of continuity implies a discontinuity in the parameter set, and these papers develop methods for simulating from the distribution of the test statistic to pin down critical values. This literature suffers from the same failure of uniform validity that plagues all "data-snooping" or adaptive estimators [66], as we document below.

With respect to this literature, the novelty of our approach is threefold: first, we are able to simultaneously test for n-th order discontinuities (points, jumps, kinks, etc). Second, we allow for multiple unknown discontinuities. There is a conventional wisdom that one could simply sequentially apply existing methods to identify multiple breaks, but this is false: the order statistic solution to the MCP is only valid for simultaneous, not sequential MCP. Moreover, as we show in Section *, even when the parametric form of the continuous part of the relationship is known and employed by the econometrician, sequential application of existing methods will often fail in the presence of multiple breaks. We emphasize this to highlight the significance of being nonparametric in the treatment of that form: even if one knows the true functional form of the continuous part, in searching for the first break the model is interim-misspecified, which can lead to erroneous results. Third and finally, however, our method is tractable and transparent.

2.2 Model

We decompose the relationship between two variables, y and x, into a continuous part and a finite set of failures of continuity. Let x be drawn from a continuous distribution with support $[\underline{\omega}, \overline{\omega}]$, a compact subset of $\mathbb{R}$. Moreover,

$$y = g(x) + \sum_{s=1...S} \sum_{k=0,...,K} d_{sk}(x) + \epsilon. \tag{2.1}$$

In this setting $g(x)$ is bounded, continuous, and differentiable up to some order n, while $\{d_{sk}(x)\}$ is a set of violations of continuity or differentiability and ϵ is an error term. With respect to notation, s indexes the location of the violation and

k indexes the degree. In particular,

$$d_{sk}(x) = \begin{cases} \psi_{s0}\mathbb{1}(x = z_s) & \text{(point discontinuity)} \\ \psi_{s1}\mathbb{1}(x \geq z_s) & \text{(jump discontinuity)} \\ \psi_{s2}\mathbb{1}(x \geq z_s)(x - z_s) & \text{(discontinuous first derivative)} \\ \psi_{s3}\mathbb{1}(x \geq z_s)(x - z_s)^2 & \text{(discontinuous second derivative)} \\ \vdots & \vdots \end{cases} \tag{2.2}$$

Here, we denote z_s as the location of the break point and ψ_s as the magnitude of the breaks. We assume no knowledge on the magnitude of breaks, the number of breaks or the type of breaks.

In empirical applications these discontinuities z often have particular economic meaning:

Example 1: Point discontinuities in [6]. In that paper, y is an expected bargaining outcome (e.g., price or first buyer offer) and x is the asking price in an online bargaining market. Round Number values of x are a signal of sellers' bargaining types, and therefore elicit discontinuously different behavior by prospective buyers. This is represented in a point discontinuity in $\mathbb{E}[y|x]$.

Example 2: Jump discontinuities in [60]. This is a classic paper in the regression discontinuity design literature which uses jump discontinuities to measure of causal treatment effects when treatment is determined by a threshold rule. In the application of [60], x is the Democrat vote share normalized to a margin of victory (or loss, for $x < 0$), y is the Democrat vote share in the subsequent election, and the paper is interested in jump discontinuities at $x = 0$ in order to study the effect of incumbency.

Example 3: Jump discontinuities in [24] [24] studies discontinuities in the dynamics of neighborhood racial composition in order to study tipping models of segregation. Here x is the current fraction of white residents, y is the rate of change of white residents, and the existence and location of a *single* jump discontinuity is unknown.

Our model is general, and its arguments may or may not have causal interpretations. For example, consider a RDD setting as in [60]. There, β_s, the parameter on a jump discontinuity at the threshold, does identify a local, causal treatment effect. However, it is *not* true that $g(\cdot)$ is a valid estimate of the counterfactual, untreated outcome except locally at z_s. Away from that point, it confounds both unobserved heterogeneity which may be correlated with x as well as heterogeneity of the treatment effect at points $x > z_s$. There is, therefore, no structural interpretation for $g(x)$ when $x \neq z_s$.

Table 2.1: coverage rate of nominal 90% ci for β

	β				
	0.00	-0.01	-0.02	-0.04	-0.16
$\hat{\beta} \pm 1.645 s(\hat{\beta})$	0.8250	0.7950	0.8050	0.8100	0.8600
Percentile	0.8100	0.8250	0.8150	0.8450	0.9200
Inverse Percentile	0.8500	0.8650	0.8700	0.9000	0.8750
Symmetric Percentile	0.8600	0.8650	0.8650	0.8950	0.8950
lasso uniform	0.8950	0.9000	0.9000	0.8950	0.9000

2.3 Assumptions and Setup

2.3.1 Assumptions

For notational purposes, let i index the observations in increasing order of x. Consistent with our assumption of a continuous distribution of x, $x_0 < x_1 < \ldots$ strictly. Note moreover that the assumption of a continuous distribution for x implies that point discontinuities are unidentified– in section 2.6.1 we will augment the model to include mass points, but for now this means we will focus on $k \geq 1$.

Even still, detection in the generic class of possible discontinuities with $k \geq 1$ remains problematic with finite data. For example, consider a jump discontinuity of known size β, and two possible locations, z and z'. If $z \leq x_i < x_{i+1} \leq z'$, then there is some hope; however, if $x_{i-1} < z < z' < x_i$, then the two are empirically indistinguishable. Therefore, for any finite sample $\{X, Y\}$, we construct the maximal meaningful set of discontinuities to be $\mathcal{Z} = \{x_i\}_{i=2,\ldots,N}$

This set of potential discontinuities is $O(N)$, and strictly larger than N if the econometrician is interested in multiple types (point, jump, kink, etc.). We cannot, therefore, simply add those discontinuities as regressors and obtain consistent estimates of their size and location using OLS. Instead, we adopt a model selection approach, and to that end we require two strong assumptions:

Assumption F. *(Sparseness) The number of discontinuities does not grow too fast, in particular* $|\{d_{sk}(x) : \psi_{sk} > 0\}|$ *is* $o(n/\log(n))$.

The stronger form of this assumption, that the number of discontinuities is

finite, is reasonable for most datasets given the bounded support of x. Sparseness is typically to be motivated by the economic intuition for the existence of the breaks in the first place, e.g. signaling conventions or institutional rules. However, when the economist is agnostic as the to existence or character of the discontinuities, as in the application to placebo tests for RD design, we acknowledge even the $o(n/\log(n))$ form of this assumption can be quite strong.

This is a parametric restriction that we require in order to implement the FDR control procedure of [46]. It is useful because it induces a exponential distribution for the p-values of the covariance test statistic for interim-significance in their forward stop algorithm, described below. Two brief notes on generality: first, none of our extensions of the [46] result employ normality, [86] derives the asymptotic of selective inference without it. The assumptions that required in [86] on error term is finite third moments. This will be guaranteed in assumption 2. Second, as an extension we relax the implied assumption of homoskedacticity in section 2.6.2 below.

We also assume $g(x)$ satisfies smoothness requirements that allows it to be approximated using nonparametric method. These assumptions are the same as those required in [36] and [49]

Assumption G. *(Smoothness)*

- *The support X is a Cartesian product of compact connected intervals on which the density $f(x)$ is bounded away from zero.*
- *$S_m(x)$ is either a spline or power series, and is nested.*
- *$g(x)$ has s times continuous derivatives on X, with $s > q/2$ for a spline and $s > q$ for a power series, where $q = dim(X)$.*

- *Define* $Q_m = E(S'_{mi}S_{mi})$. *There exist* $\alpha > 0$, $\eta > 0$, $\phi < \infty$, *such that for all* $l'Q_m l = 1$ *and* $0 \le \mu \le \eta$, $\sup_m \mathbb{P}(|l'z_{mi}| \le u) \le \phi\mu^\alpha$
- $\max_{m<M_n} T_m^4/n = O(1)$ *for a power series or* $\max_{m<M_n} T_m^3/n = O(1)$ *for splines sieve.*
- $\phi_m^2 < 0$ *for all* $m < \infty$
- *For some* $N > 0$,
  1. $\sup_i \mathbb{E}(e_i^{4(N+1)}|x_i) < \infty$
  2. *Let* $q_{jn} = \#\{m : T_m = j\}$ *be the number of models which have exactly* j *coefficients, and* $\bar{q}_n = \max_{j\le M_n} q_{jn}$. $\bar{q}_n = o(\xi_n^{1/N})$, *where* $\xi_n = \inf_m nIMSE_n^*(m)$
  3. *Define* $h_{mi} = S'_{mi}(\sum_{i=1}^n S_{mi}S'_{mi})^{-1}S_{mi}$. *Then* $\max_{m\le M_n}\max_{i\le n} h_{mi} \to 0$ *almost surely*

2.3.2 Notation and Setup

For a given finite sample $\{(x_i, y_i)\}_{i=1}^n$, let $(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$ as the ordered statistic of $\{x_i\}_{i=1}^n$. Define $D(x_i)$ as

$$\begin{cases} D(x_i) = \left(1_{x_i=x_{(1)}}, 1_{x_i=x_{(2)}}, \cdots, 1_{x_i=x_{(n)}}\right) & \text{(point discontinuity)} \\ D(x_i) = \left(\frac{1_{x_i>x_{(1)}}}{\sqrt{n-1}}, \frac{1_{x_i>x_{(2)}}}{\sqrt{n-2}}, \cdots, \frac{1_{x_i>x_{(n-1)}}}{1}\right) & \text{(jump discontinuity)} \\ D(x_i) = \left(\frac{(x_i-x_{(1)})\times 1_{x_i>x_{(1)}}}{\phi_1}, \frac{(x_i-x_{(2)})\times 1_{x_i>x_{(2)}}}{\phi_2}, \cdots, \frac{(x_i-x_{(n)})\times 1_{x_i>x_{(n-1)}}}{\phi_{n-1}}\right) & \text{(kink discontinuity)} \end{cases}$$

where $\phi_k = \sqrt{\sum_{i=k+1}^n (x_{(i)} - x_{(k)})^2}$. We normalize all the regressors to have norm 1 so they are balanced when penalized by lasso.

We also allow those cases when different types of discontinuities exist. For

examle when both kink and jump exists:

$$D(x_i) = \Bigg(\frac{1_{x_i > x_{(1)}}}{\sqrt{n-1}}, \frac{1_{x_i > x_{(2)}}}{\sqrt{n-2}}, \cdots, \frac{1_{x_i > x_{(n-1)}}}{1}, \frac{(x_i - x_{(1)}) \times 1_{x_i > x_{(1)}}}{\phi_1},$$

$$\frac{(x_i - x_{(2)}) \times 1_{x_i > x_{(2)}}}{\phi_2}, \cdots, \frac{(x_i - x_{(n)}) \times 1_{x_i > x_{(n-1)}}}{\phi_{n-1}} \Bigg)$$

Define $S(x_i)$ as the SIEVE expansion regressor vectors. And let λ be the lasso tunning parameter.

Our procedure concerns itself with the following LASSO regression:

$$(\tilde{\beta}, \tilde{\psi}) = \arg\min_{\beta,\psi} \mathbb{E}(y_i - S(x_i)\beta - D(x_i)\psi)^2 + \lambda|\psi|_1 \tag{2.3}$$

– where $S(x_i)$ is a vector of linear basis functions for the space of continuous functions on $[\underline{\omega}, \overline{\omega}]$, e.g. basis splines, and $D(x_i)$ is the set of potential discontinuities of interest. Note that this LASSO specification is unique in that β is entirely unpenalized– this is the sense in which our approach maintains the null of continuity. Our regression function biases the estimator in favor of representing the data using a continuous function.[1]

The critical step is in the choice of λ. The standard approach would be to choose λ by either cross-validation or by using an estimate of σ_ϵ to approximate the rate-optimal λ. While convenient, these leave the researcher neither an inferential framework nor control over the Type I error rate– in our setting, the likelihood of falsely detecting discontinuities. Instead, we employ the FDR LASSO framework of [46] in order to control the probability of false inclusion of

[1]This representation was first introduced in the working paper version of [6], however that paper only displayed coefficient paths as motivation for model selection: it not develop the inferential framework for interpreting the results of the regression.

variables in the model. In particular, we use their *forward stop* algorithm, which proceeds path-wise along a sequence of covariance test statistics that offer interim significance tests for the marginal included variable [69].

[Add more discussion of the [46] method; for now, see their paper]

Two few notes about the character of the test: First, it should not be surprising that it exhibits variable power, depending on the quality of the local approximation to $g(x)$. In regions where there is little data, our approach will fail to detect discontinuities. While it is tempting to interpret this as a bias in favor of flagging discontinuities where data is abundant, say near x', rather than where it is not, say near x'', this is not precisely correct: it is a question of variable power. Rather, one might say that it is a bias in favor of flagging discontinuities near x' rather than x'' *conditional on the existence of discontinuities near both*.

Second, as in the derivation of the rate-optimal λ, our approach requires an estimate of σ_ϵ at every stage of the forward-stop algorithm. We estimate this object using OLS conditional on the discontinuities included so far, so that the estimate is consistent under the interim null of the sequential MCP.

2.3.3 Irrepresentable Condition

Consider event B. First notice that fix $k_0 = 0$, ($A_0 = \emptyset$, $s_{A_0} = \emptyset$), $\mathbb{P}(B) = 1$ holds trivially.

A necessary and sufficient condition for $\mathbb{P}(B) \to 1$ is the irrepresentable condition in [93],

Definition 1. *We say that the irrepresentabile condition holds for $\eta < 1$, if*

$$\max_{j \notin A_0} \sup_{\|\tau_{A_0}\|_\infty \leq 1} \left| D_j' D_{A_0} (D_{A_0}' D_{A_0})^{-} \tau_{A_0} \right| < \eta$$

In our jump discontinuity design matrix, D is a lower triangular matrix after rearranging the rows. For example, the kth column is:

$$D_k = \left(0, 0, 0, \cdots, \frac{1}{\sqrt{n-k}}, \frac{1}{\sqrt{n-k}}, \cdots, \frac{1}{\sqrt{n-k}}\right)'$$

Assume $A_0 = \{k\}$, that is we only have 1 single jump discontinuity. Thus $D_{A_0}' D_{A_0} = 1$ and for all $j \neq k$,

$$D_j' D_{A_0} = \frac{\min\{(n-k), (n-j)\}}{\sqrt{(n-k)(n-j)}} < 1$$

Thus the irrepresentable condition holds.

Theorem 5. *(1) For any set $A_0 \subset \{1, 2, \cdots, p\}$, the design matrix for point discontinuity satisfies the irrepresentable condition.*

(2) For any set $A_0 \subset \{1, 2, \cdots, p\}$, the design matrix for jump discontinuity satisfies the irrepresentable condition.

(3) For any set $A_0 \in \{1, 2, \cdots, p\}$ as a singleton, the design matrix for kink discontinuity satisfies the irrepresentable condition.

(4) For any set $A_0 \in \{1, 2, \cdots, p\}$ as a singleton, the design matrix for kink+jump discontinuity satisfies the irrepresentable condition.

Corollary 3. *Let P_z a projection matrix such that $P_z X_{A_0}$ has full column rank. If the design matrix X satisfies the irrepresentable condition, then $P_z X$ also satisfies the irrepresentable condition.*

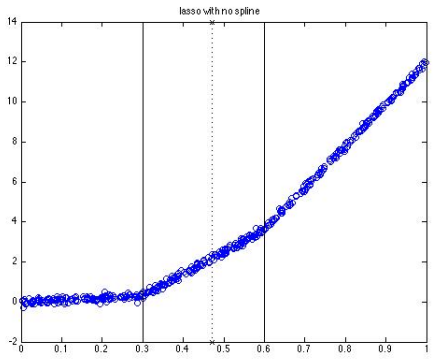
However, it is also possible to demonstrate the the irrepresentable condition will not hold for many cases of interest.

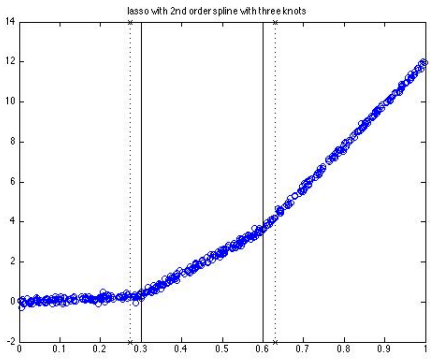
Corollary 4. *Irrepresentable condition does not hold when there are two or more kinks*

For two and more kinks, the irrepresentable condition does not hold. However, we can assume there exits a partition on the support of X such that each segment contains at most one kink discontinuity.

Then our method can be applied to each segment. Basis spline provides a nature way to combine nonparametric estimation and partition.

As illustrated in the following example, our method can identify two kinks using basis spline.





2.4 Detecting Discontinuities

2.4.1 Covariance Test

Ignore the term $S(x_i)$ in (3) and considr the standard lasso setup in matrix form

$$\hat{\Psi} = \arg\min_{\Psi} \frac{1}{2}\|Y - D\Psi\|_2^2 + \lambda\|\Psi\|_1$$

where $\psi = (\psi_1, \psi_2, \cdots, \psi_p)'$, $D = (D(x_1)', D(x_2))', \cdots, D(x_n)')'$, and denote D_j the jth column of D. The solution $\Psi(\lambda)$ is a continuous and piecewise linear function and can be computed via the LARS algorithm as [37].

Define $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ as he knot (changes in slope) on $\Psi(\lambda)$. Let $< ., . >$ be the inner product operator. Let A be the active set before the knot λ_k, and suppose predictor j enters at λ_k. Define $\Psi_A(\lambda_{k+1})$ be the lasso solution at λ_{k+1} but constraint on the set A:

$$\Psi_A(\lambda_{k+1}) = \arg\min_{\Psi_A} \frac{1}{2}\|Y - D_A\Psi_A\|_2^2 + \lambda_{k+1}\|\Psi_A\|_1$$

Then the covariance test statistic as proposed in [69] is:

$$T_k = (< y, D\Psi(\lambda_{k+1}) > - < y, D_A\Psi_A(\lambda_{k+1}) >) / \sigma^2$$

At each step k, the null hypothesis we are testing is $A \supseteq A^* = supp(\Psi^*)$, where Ψ^* is the true parameter.

Let sign vector $s_{A_0} \in \{-1, 1\}^{|A_0|}$. Consider the event:

$$B = \Big\{\text{The solution at step } k_0 \text{ in the lasso path has active set } A = A_0,$$
$$sign s_A = sign((D_{A_0})^+ y) = s_{A_0}, \text{ and the next two knots are given by,}$$
$$\lambda_{k_0+1} = \max_{j \notin A \cup \{j_{k_0}\}, s \in \{-1,1\}} \frac{D_j'(I - P_A)y}{s - D_j'(D_A')^+ s_A}, \lambda_{k_0+2} = \lambda_{k_0+2}^{join}\Big\}$$

Assume $\mathbb{P}(B) \to 1$ as $p \to \infty$, or in other word, all active variables enter the Lasso path first, [69] shows that

$$\lim_{p \to \infty} \mathbb{P}(T_k > t) \leq e^{-t}$$

Thus standard exponential distribution serves as a conservative bound.

2.4.2 Sequential False Discovery Rate Control

Testing from step 0 to n, we are dealing with a sequence of model:

$$\emptyset = M_0 \subset M_1 \subset \cdots \subset M_n, \text{ with } M_k \subset \{1, 2, \cdots, n\}$$

This process corresponds to a sequence of $p-$value $p_1, p_2, \cdots, p_n$ from the covariance test at each step.

In such a multiple comparison setting, the False Discovery Rate (FDR) is defined as $\mathbb{E}\Big[V(\hat{k})/\max(1, \hat{k})\Big]$, where $V(\hat{k})$ is the number of null hypotheses among all the rejected hypotheses.

Since our hypothesis are nested. The $\hat{k}$th hypothesis will only be rejected if the previous $\hat{k} - 1$ hypothesis are all rejected. [46] propose two stopping rules for $\hat{k}$:

- $$\hat{k}_F = \max\left\{k \in \{1, \cdots, n\} : -\frac{1}{k}\sum_{i=1}^{k} \log(1 - p_i) \leq \alpha\right\}$$
- $$\hat{k}_S = \max\left\{k \in \{1, \cdots, n\} : \exp\Big(\sum_{i=1}^{n} \log \frac{\log p_j}{j}\Big) \leq \frac{\alpha k}{n}\right\}$$

The first one is called forward stopping rule, it is moderately robust to potential misspecification of the null hypothesis.

The second one is called String stopping rule, which also control the Family-Wise Error Rate (FWER) at level α. It is more conservative than the forward stopping rule.

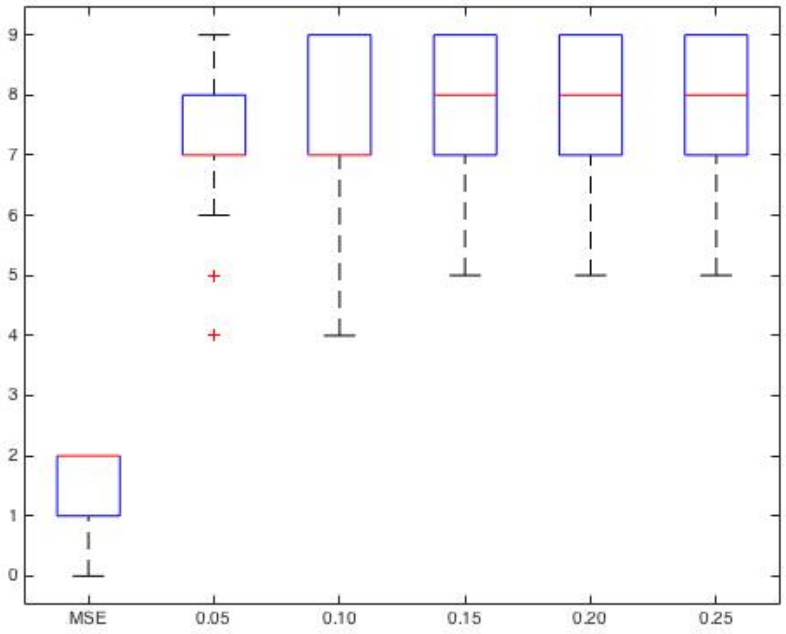
In the following application, we use covariance test and forward stopping rule to test discontinuities while controlling the FDR.

2.4.3 Advantage of Lasso

Compare with traditional mean squared error method the lasso type estimator has two main advantages.

- Efficiency in the detection
- Uniformly valid inference for the magnitude of the breaks

The traditional MSE method requires sample splitting after a break is been detected. On the other hand, lasso estimate subtracts (part of) the break from the sample and then using the entire sample to detect the next break. Thus, there is no loss in efficiency.



2.5 Asymptotic Properties

2.5.1 Consistency

Using a finite sample criterion like FDR to choose λ does not guarantee consistent variable selection that often been granted to lasso procedure. Two types of error can be made: (1) a true break is not detected (type II error); (2) a detection is not a true break (type I error). So instead, we show that even we miss some ingredients that is in the true data generating process, the coefficients or magnitude of those variables are bounded and their effect on the IMSE goes to 0 as n goes to infinity. On the other hand, a false detection will not affect the model fitting asymptotically.

Formally, let (y_i, x_i), $i = 1, \cdots, n$ be the sample we observed. We suspect the conditional mean is $g(x) = E(y|x)$ with breaks at $z_1 < z_2 < \cdots < z_p$. Let e_i be a mean 0 process with variance σ^2 unknown. The data generating process can be summarized as:

$$y_i = g(x_i) + \sum_{j=1}^{p} \psi_j \times L_j(x_i) + e_i$$

where $L_j(x_i)$ represents different kinds of discontinuities.

Let $(\hat{\beta}, \hat{\psi})$ be the estimator from (3). Define $\hat{I} = \{i, \hat{\psi}_i \neq 0\}$. Define $D_l(x_i)$ as the lth term in $D(x_i)$. The integrated mean squared error (IMSE) for our estimator can be written as:

$$IMSE_n(m) = \int \mathbb{E}\left(\hat{g}_m(x) + \sum_{l \in I} D_l(x)\hat{\psi}_l - g(x) - \sum_{j=1}^{p} L_j(x)\psi_j \right)^2 f(x)dx$$

where

$$\hat{g}_m(x) + \sum_{l \in I} D_l(x)\hat{\psi}_l = S_m(x)'\hat{\beta}_m + \sum_{l \in I} D_l(x)\hat{\psi}_l$$
$$g(x) + \sum_{j=1}^{p} L_j(x)\psi_j = S_m(x)'\beta_m + \sum_{j=1}^{p} L_j(x)\psi_j + r_m(x)$$

we first show that minimization problem (3) is equivalent to a standard lasso problem.

Lemma 1. *The problem of (3) is equivalent to the following standard LASSO problem:*

$$(\tilde{\psi}) = \arg\min_{\psi} ||MY - MD\psi||_2^2 + \lambda|\psi|_1$$

– where M is the projection matrix $I_{n\times n} - S_m(S_m'S_m)^{-1}S_m$.

Theorem 6. *Let ω_m be the compatibility constant for the sequence of design matrices M_mD_n. If there exists ω such that for all $m > m_0$, $|\omega_m| > \omega$*

$$IMSE_n(m) \leq 2\phi_m^2 + 2\sigma^2\frac{K_m}{n} + 8W_1^2\sigma^2\frac{\log p_n}{n}s_0/\omega^2$$

2.5.2 Distribution of Break Point

We compare our estimator with Bai 1993 and consider the limiting distribution of the location of the break when the magnitude of the break is shrinking to 0 as n goes to infinite. We derive the distribution of the estimated break point under the assumption that only one jump (or kink) break exists. We show that the lasso detected location has similar distribution as the traditional mean square error estimates.

Theorem 7. *Assume there is only one jump break such that the data generating process is:*

$$y_i = L_1^0(x_i)\psi_1 + \epsilon_i,$$

$$L_1(x_i) = \begin{cases} 0 & \text{if } x_i \leq z \\ 1 & \text{if } x_i > z \end{cases}$$

Let $k > 0$ be an integer such that

$$x_{(k)} \leq z < x_{(k+1)}$$

Let $\rho_n = O_p(\|\psi_1\|)$ and $\hat{k}$ be the lasso estimator from (3). Assume $\rho_n \to 0$ but $\sqrt{n}\rho_n \to \infty$. There exist a constant v, such that

$$\rho_n^2(\hat{k} - k) \to_d \arg\max_v(-|v| + 2W(|v|))$$

where $W(v)$ is a wiener process of degree v

For kink, we derive the following theorem

Theorem 8. *Assume there is only one kink break such that the data generating process is:*

$$y_i = L_1^1(x_i)\psi_1 + \epsilon_i,$$

$$L_1^1(x_i) = \begin{cases} 0 & \text{if } x_i \leq z \\ (x_i - z) & \text{if } x_i > z \end{cases}$$

Let $k > 0$ be an integer such that

$$x_{(k)} \leq z < x_{(k+1)}$$

Let $\rho_n = O_p(||\psi_1||)$ and $\hat{k}$ be the lasso estimator from (3). Assume $\rho_n \to 0$ but $\sqrt{n}\rho_n \to \infty$. Assume x is has compact support $[x^-, x^+]$. There exist constant $v_1 \leq 0$ and $v_2 > 0$, such that

$$\rho_n^2(\hat{k} - k) \to_d \begin{cases} \arg\max_v(v_1 T_1 + 2W(-v_1)) & \hat{k} \leq k \\ \arg\max_v(-v_2 T_2 + 2W(v_2)) & \hat{k} > k \end{cases}$$

where $W(v)$ is a wiener process of degree v. $T_1 = \int_{x^-}^{x_{(k)}} (u - x^-)^2 f(u)du$ and $T_2 = \int_{x_{(k)}}^{x^+} (u - x^+)^2 f(u)du$

2.5.3 Uniformly Valid Inference for Magnitude of Breaks.

To perform inference on ψs when the location is unknown, one concern is the uniformity on the post model selection estimator. For example, after the location of the break point $\hat{k}$ is estimated, we can fit the model using the $\hat{k}$th regressor and estimate ψ.

These kind of estimates are not uniformly consistent as shown in (Leeb and Postcher 2008). More precisely, the finite sample behavior of those estimates are not properly reflected by their point-wise limit. In our method, we propose a uniformly valid inference for our lasso estimates.

Consider the simply model:

$$y_i = d_{1,i}\psi_1 + d_{2,i}\psi_2 + \alpha + \epsilon_i$$

If we impose the sparsity assumption or there are no more than s breaks in the data and consider the set of parameters:

$$B(s) = \{\psi \in R^2 | \{j, \psi_j \neq 0\} \leq s\}$$

Assume no more than one of the ψs are non-zero, we are interested in constructing a uniform consistent estimator $\tilde{\psi}$, such that

$$\sup_{\psi_0 \in B(1)} P\left(\left|(\tilde{\psi}_j - \psi_{0,j}) > \delta\right|\right) \to 0 \tag{2.4}$$

The possible true model space given the sparsity assumption are:

$$M_0 = \begin{cases} R_1 \text{ if } \psi_1 = 0, \psi_2 \neq 0 \\ R_2 \text{ if } \psi_2 = 0, \psi_1 \neq 0 \\ R_3 \text{ if } \psi_2 = 0, \psi_1 = 0 \end{cases}$$

Given the model selection procedure, we have the unconditional distribution for ψ_1 and ψ_2 as

$$\tilde{\psi}_1 = 0 \cdot 1(\hat{M} = R_1) + \hat{\psi}_1(R_2) \cdot 1(\hat{M} = R_2) + 0 \cdot 1(\hat{M} = R_3)$$

$$\tilde{\psi}_2 = \hat{\psi}_2(R_1) \cdot 1(\hat{M} = R_1) + 0 \cdot 1(\hat{M} = R_2) + 0 \cdot 1(\hat{M} = R_3)$$

Notice that function $1(\hat{M} = R_2)$ will depends on the true value of both ψ_1 and ψ_2. When $\psi_{0,1}$ is approaching 0 at rate $\frac{1}{\sqrt{n}}$, (4) can be violated as shown in Leeb and Postcher.

We propose the use of de-bias lasso from (Van de Geer 2014). The intuition behind is we drop the sparsity assumption when constructing confidence intervals and thus avoid working on the non-convex set $B(s)$.

2.6 Extensions

2.6.1 Point Discontinuities

As noted above, point discontinuities are not identified when x is continuously distributed. To accommodate point discontinuities, as in the [6] example, one could simply assume that x is distributed according to a mixture of a continuous distribution and a discrete distribution with finitely many points of support. Then one can include point discontinuities as in (2.2) above.

Under this framework our asymptotic results for IMSE will still hold, however note that in finite samples these point discontinuities will not be identified as against a pair of symmetric jump discontinuities, one positive and one negative, at exactly that point. Such jump discontinuities would incur twice the penalty as a point discontinuity, and in this sense the LASSO preference for parsimony will cause it to select the point discontinuity.

2.6.2 Heteroskedacticity

Under the presence of Heteroskedasticity, theorem 1 in Meinshausen and Yu (2009) is still valid.

Formally, let $e_i \sim N(0,1)$. The data generating process can be summarized as:

$$y_i = g(x_i) + \sum_{j=1}^{p} \psi_j L_j(x_i) + \sigma(x_i) e_i \tag{2.5}$$

$\Sigma = diag(\sigma(x_1), \sigma(x_2), \cdots, \sigma(x_n))$ is a positive definite matrix.

We propose the following post selection estimator:

- First solve the following penalized estimators:

$$(\tilde{\beta}, \tilde{\psi}) = \arg\min_{\beta,\psi} \mathbb{E}\left(\frac{y_i - s'_{mi}\beta - D_i\psi}{\sigma(x_i)}\right)^2 + \lambda|\psi|_1 \tag{2.6}$$

  where

  $s_{mi} = S_m(x_i)$ are the SIEVE expansion regressor vectors.

  D_i is defined as previous based on the ordered statistic.

  λ is the turning parameter of lasso

- Estimate the post selection model:

$$(\hat{\beta}, \hat{\psi}) = \arg\min_{\beta,\psi} \mathbb{E}\left(\frac{y_i - s'_{mi}\beta - \hat{D}_{Ii}\psi}{\sigma(x_i)}\right)^2 \tag{2.7}$$

  where

  I is the non-zero terms for $\tilde{\psi}$ selected at first stage.

  $\hat{D}_{Ii}$ is D_i restricted to I.

This estimator is similar to the Wagener and Dette (2011)'s weighted penalized least squares estimator. However, their estimator only address for fixed p cases. With the previous proof, the estimator above could be used for $p > n$ case.

Write $Y = S\beta + D\psi + \Sigma e$, then

$$Y = S\beta + D\psi + \Sigma e \Leftrightarrow \Sigma^{-1}Y = \Sigma^{-1}S\beta + \Sigma^{-1}D\psi + e \tag{2.8}$$

For an estimator $\hat{\Sigma}$ of Σ, consider the difference of the minimization target:

$$L(\beta,\psi,\sigma) - L(\beta,\psi,\hat{\sigma}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - s_i'\beta - D_i\psi}{\hat{\sigma}(x_i)}\right) + \lambda|\psi|_1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - s_i'\beta - D_i\psi}{\sigma(x_i)}\right) - \lambda|\psi|_1$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - s_i'\beta - D_i\psi)(\frac{1}{\hat{\sigma}(x_i)} - \frac{1}{\sigma(x_i)})\right) \tag{2.9}$$

So as long as $\hat{\sigma}(x_i) \to \sigma(x_i)$, we can construct the uniform consistency between $(\hat{\beta}_1, \hat{\psi}_1) = \arg\min(L(\beta,\psi,\sigma))$ and $(\hat{\beta}_2, \hat{\psi}_2) = \arg\min(L(\beta,\psi,\hat{\sigma}))$

I propose the following estimator for $\hat{\sigma}(x_i)$:

- solve the following penalized estimators, using cross-validation:

$$(\tilde{\beta}_1, \tilde{\psi}_1) = \arg\min_{\beta,\gamma} \mathbb{E}\left(y_i - s_{mi}'\beta - D_i\psi\right)^2 + \lambda|\psi|_1 \tag{2.10}$$

- Calculate $\hat{\epsilon}_i = y_i - s_{mi}'\tilde{\beta}_1 - D_i\tilde{\psi}_1$
- Compute local linear fit of squared residual in order to estimate the conditional variance $\hat{\sigma}(x_i)$

One thing remains to justify is the behavior of lasso with heteroskedasticity, like (6).

Assumption H. *The function $|\sigma(x_i)|$ is uniformly bounded by some M_σ*

Corollary 5. *Assume $W_\sigma < \infty$ and assume $\sigma(x)$ has bounded second derivatives on X. The density $f(x)$ is a Lipschtiz function. If there exists ω such that for all $m > m_0$, $|\omega_m| > \omega$. Then*

$$IMSE_n(m) \leq 2\phi_m^2 + 2W_\sigma^2\frac{K_m}{n} + 8W_1^2W_\sigma^2\frac{\log p_n}{n}s_0/\omega^2$$

2.7 Applications

2.7.1 Placebo Tests for Structural Breaks

A natural application of our procedure is to placebo testing for discontinuities in the RDD setting. While the location of a break is often known in advance from institutional details — for instance, majority rule voting implies a threshold at fifty percent — empirical researchers in this area would like to validate the existence of these discontinuities and know whether there are other, unanticipated ones that may affect results. This takes advantage of the unique feature of our method which is that it is valid in the presence of multiple breaks.

By way of illustration, we apply our method to the electoral RD design of [60]. Happily we detect the discontinuity at the vote share margin of winning of zero (to be precise, 0.0003), and we detect no further discontinuities, as depicted in Figure 1.

2.7.2 Test discontinuity with unknown location

We follow [24] and use our method to test race-based tipping in neighborhoods. Let m_t denotes the minority share at time t. Their theory suggested that

$$\mathbb{E}(\Delta m_t | m_{t-1}) = \mathbf{1}(m_{t-1} < m^*) g(m_{t-1}) + \mathbf{1}(m_{t-1} \geq m^*) h(m_{t-1}),$$

for some threshold m^* and some functions g and h.

We use the Neighborhood Change Database (NCDB) from 1970-2010. The change in minority shares are calculated using ten-year window. All settings are

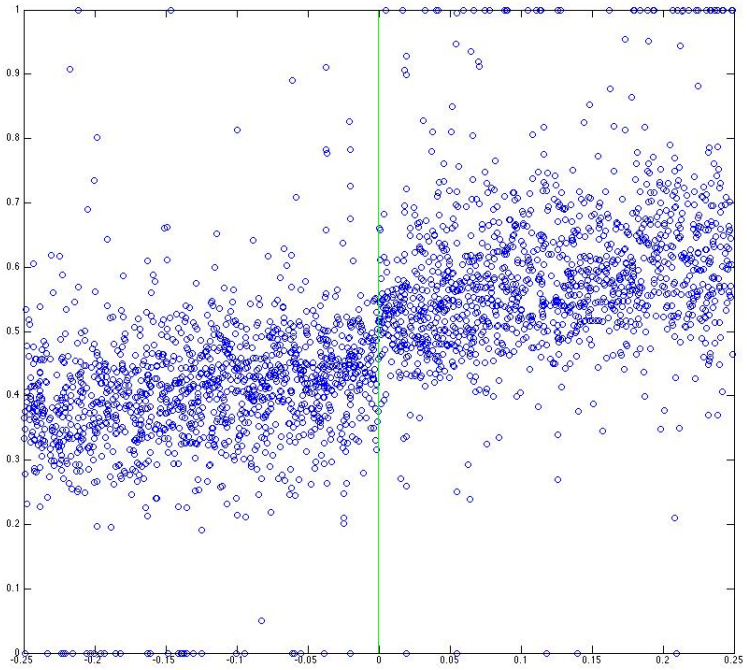


Figure 2.1: Lee 2008 data

exactly the same as in [24]. In the figure below, we plot the change in the tract-level non-Hispanic white population as a percentage of the total tract population. A vertical dash line represents a detected discontinuity and the horizontal line represents the unconditional mean of the data. We plot the data for major metropolitan areas: Chicago, San Jose and New York. We find tipping points in Chicago and New York through out our testing periods. We find tipping points in San Jose through 1970-1990 period but not through 1990-2010. We compare our finding with census segregation map and find our results consistent with the pattern on the maps.

2.7.3 Regression Kink with an Unknown Threshold

One may wonder how our procedure compared to traditional structure break literature. We compare our method to Bruce Hansen's 2015 paper and show that we achieve a similar size and power in the detection of the first discontinuity.

More specifically, we follow the same simulation example in Hansen 2015:

X is the debt/GDP ratio from 1792-2009 as in [81]. The data generating process is as following:

$$y_t = \beta_1(x_t - \gamma)_- + \beta_2(x_t - \gamma)_+ + \beta_3 y_{t-1} + \beta_4 + \epsilon_t \tag{2.11}$$

where $\epsilon_t \sim N(0, \sigma^2)$.

To evaluate size, we set $\beta_1 = \beta_2 = 0, \beta_3 = 0.3, \beta_4 = 3$, and $\sigma^2 = 16$ to match the empirical estimates.

We report 1000 simulation. The bootstrap size for Hansen 2015 is set at 1000. Both tests exhibit no meaningful size distortion.

Table 2.2: size

	α				
	0.05	0.10	0.15	0.20	0.25
Hansen	0.051	0.0990	0.1469	0.1970	0.2460
Lasso	0.071	0.1120	0.1520	0.2040	0.2450

Second, we compare the power of the two tests. Change the value of β_2 from -0.16 to -0.02 and set the kink point at $\gamma = 40$. Nominal size is set at 0.05.

Table 2.3: power

	β							
	-0.02	-0.04	-0.06	-0.08	-0.10	-0.12	-0.14	-0.16
Hansen	0.0691	0.1130	0.2010	0.3110	0.4680	0.6200	0.7500	0.8540
Lasso	0.0920	0.1540	0.2626	0.3740	0.4690	0.5460	0.6690	0.6890

Table 2.4: bootstrap

	α				
	0.05	0.10	0.15	0.20	0.25
Hesen 2015:					
$\hat{\gamma} \pm 1.645 s(\hat{\gamma})$	0.7900	0.7200	0.6850	0.6450	0.6050
	(37.9457)	(31.8450)	(27.8699)	(24.8114)	(22.2712)
Percentile	0.9550	0.8900	0.8650	0.8000	0.7650
	(40.7975)	(32.5250)	(27.0200)	(23.2675)	(20.0700)
Inverse Percentile	0.8250	0.7950	0.7350	0.7100	0.6800
	(40.7975)	(32.5250)	(27.0200)	(23.2675)	(20.0700)
Symmetric Percentile	0.9000	0.8450	0.7800	0.7500	0.7150
	(41.9400)	(32.1950)	(26.5050)	(22.5850)	(19.5100)
C_γ	0.9050	0.8400	0.7800	0.7150	0.6450
	(30.6950)	(24.6100)	(20.9950)	(18.4050)	(16.0800)
C_{γ^*}	0.9200	0.8450	0.8050	0.7700	0.7250
	(33.3950)	(27.3200)	(23.5250)	(20.9050)	(18.6600)
Lasso:					
Percentile	0.9500	0.8650	0.8000	0.7700	0.7250
	(28.0704)	(22.3397)	(18.9569)	(16.6049)	(14.5919)
Inverse Percentile	0.9200	0.8550	0.7850	0.7350	0.6850
	(28.0704)	(22.3397)	(18.9569)	(16.6049)	(14.5919)
Symmetric Percentile	0.9500	0.8600	0.7950	0.7550	0.6900
	(28.0731)	(22.2410)	(18.9754)	(16.4391)	(14.4893)

CHAPTER 3

LOCAL REGRESSION SMOOTHERS WITH SET-VALUED OUTCOME DATA

3.1 Introduction

Statistical analysis has traditionally contended with problems of data imprecision due to limits in the measuring instruments and to measurement error, as well as with missing data, data coarsening and grouping. In other words, it has contended with different forms of set-valued data. Within the social sciences in particular, collection of data in the form of sets, especially intervals, has become increasingly widespread. For example, the Health and Retirement Study is one of the first surveys where, in order to reduce item non-response, income data is collected from respondents in the form of brackets, with degenerate (singleton) intervals for individuals who opt to fully report their income (see, e.g. [54]). To reduce response burden, the Occupational Employment Statistics (OES) program at the Bureau of Labor Statistics collects wage data from employers as intervals, and uses these data to construct estimates for wage and salary workers in 22 major occupational groups and 801 detailed occupations. Privacy concerns often motivate providing public use tax data as the number of tax payers which belong to each of a finite number of cells. In the medical field, due to ethical and cost reasons, time-to-event measurements are not collected on a continuous scale, but at pre-specified time intervals. And many more examples are possible.

The partial identification literature in econometrics (e.g., [75]) has addressed the question of what can be learned about functionals of probability distributions of interest, when some of the variables are only known to belong to (ran-

dom) sets and no assumptions are imposed on the distribution of the true variables within these sets. We take the identification results of this literature as our point of departure. Our contribution is to provide statistical results on local linear regression smoothing when the outcome data is set valued and the regressors are exactly measured.

Specifically, the paper relaxes the textbook setting (e.g., [87]) of non-parametric regression –where regressors and outcome data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, are precisely measured– by assuming that $\boldsymbol{y}_i$ is only known to belong to an observed set $\boldsymbol{Y}_i$. In other words, we deal with an independently and identically distributed sample of observations for the pair $(\boldsymbol{x}, \boldsymbol{Y})$ composed of a random vector $\boldsymbol{x}$ belonging to an hyper-rectangle $I^m \subset \mathbb{R}^m$ and a random convex compact set $\boldsymbol{Y}$ in $\mathbb{R}^d$. Here $\boldsymbol{Y}$ is assumed measurable in a sense made precise in Section 3.2. The true (however unobservable) outcome associated with $\boldsymbol{x}$ is a random vector $\boldsymbol{y}$ that almost surely takes values in $\boldsymbol{Y}$.

Our goal is to provide a non-parametric regression estimator for the expectation conditional on $\boldsymbol{x}$ of each random vector $\boldsymbol{y} \in \boldsymbol{Y}$. For a given tuple $(\boldsymbol{x}, \boldsymbol{y})$ that almost surely belongs to $\{\boldsymbol{x}\} \times \boldsymbol{Y}$, we denote by

$$m(x) = \mathbb{E}[\boldsymbol{y} | \boldsymbol{x} = x]$$

the regression function for the chosen $(\boldsymbol{x}, \boldsymbol{y})$. Each choice of $(\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s. gives rise to a function m and we denote by $\mathcal{M}$ the family of all regression functions generated in this manner. Additionally, we denote

$$M(x) = \{m(x) : m \in \mathcal{M}\},$$

and we observe that

$$M(x) = \mathbb{E}[\boldsymbol{Y} | \boldsymbol{x} = x] = \left\{\mathbb{E}[\boldsymbol{y} | \boldsymbol{x} = x] : \boldsymbol{y} \in \boldsymbol{Y} \text{ a.s.}\right\}$$

is the conditional selection expectation of $\boldsymbol{Y}$, see [78, Section 2.1.6] and Section 3.2 below. For example, consider the empirically relevant case that $d = 1$ and $\boldsymbol{Y} = [\boldsymbol{y}_L, \boldsymbol{y}_U]$ for two random variables $\boldsymbol{y}_L, \boldsymbol{y}_U$ such that $\mathbf{P}(\boldsymbol{y}_L \leq \boldsymbol{y}_U) = 1$. Then

$$M(x) = \Big[\mathbb{E}[\boldsymbol{y}_L|\boldsymbol{x} = x], \mathbb{E}[\boldsymbol{y}_U|\boldsymbol{x} = x]\Big]. \tag{3.1}$$

Our proposal is to estimate $M(x)$ as a weighted sum of the sets $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ using local linear estimator.[1] The development of our technical results directly builds on classic references such as [39] and [38], and is closely related to [40] and [87]. Inspection of equation (3.1) might suggest, for the case $d = 1$, to report an estimator given by the interval between a local constant or local linear regression of $\boldsymbol{y}_L$ on $\boldsymbol{x}$ and of $\boldsymbol{y}_U$ on $\boldsymbol{x}$. While both in finite sample and asymptotically this approach is equivalent to what we propose for the case of a local constant regression, for the case of local linear regression equivalence breaks down in finite sample and affects the bias in the asymptotical setting. The difference is important: we show in Remark 9 below that reporting the interval between a local linear regression of $\boldsymbol{y}_L$ on $\boldsymbol{x}$ and of $\boldsymbol{y}_U$ on $\boldsymbol{x}$ may lead to a finite sample bias understating the width of $M(x)$ and is therefore quite unpalatable. For example, such estimator might be empty or a singleton in finite sample even though $M(x)$ is an interval of strictly positive width in population. In contrast, the estimator that we propose does not suffer from this problem, although it does have an asymptotic bias term similar to that of point identified local linear regression estimators.

We derive the asymptotic properties of our estimator, propose a bias correction method, and adapt results from [15] to obtain pointwise confidence bands

[1] We comment on the case of local constant (Nadaraya–Watson) estimator and discuss the difficulties associated with generalizing our method to local polynomial estimators of order larger than one in our concluding section.

that asymptotically cover the functional of interest with probability $1-\alpha$. We report the results of Monte Carlo simulations with interval valued $\boldsymbol{Y}$ that support our theoretical findings.

We also demonstrate the usefulness of our approach with an empirical illustration. We use a novel dataset that follows 132 patients during anti-cancer treatment regimens between March 11, 2010 and April 29, 2016. These patients are affected by advanced non-squamous non small cell lung cancer and present an epidermal growth factor receptor (EGFR) of the wild type. Interest centers on estimating the expected time from patient registration until tumor progression (TTP) conditional on a measure of gene CDC25A at baseline. Tumor progression is defined as an increase in the diameter of the tumor of 20% compared with the smallest diameters of all previous tumor assessments (this is called RECIST criterion in the medical literature). CDC25A is a gene involved in cell division, and it is considered an oncogene playing a role in DNA damage control. The oncogene status of CDC25A suggests that its overexpression is associated with a poorer prognosis with regard to its biological role. It is therefore of interest to immunologists to quantify the relationship between TTP and CDC25A. However, tumor progression time can only be measured by intervals, and no information is available on the distribution of true TTP within the interval. Our method provides a consistent estimator of the set of admissible values for the conditional expectation of interest, as well as $1-\alpha$ pointwise confidence bands for it. Our results suggest that expected TTP is decreasing in the expression of gene CDC25A.

Related literature. Within the partial identification literature, there is a large body of work analyzing regression with interval valued data. [73] consider

models where one variable (either outcome or covariate) is observed as intervals and all others are perfectly measured, and provide identification results for nonparametric as well as parametric models in this setting. [15] introduce to the partial identification literature the use of random set theory and provide results on identification and inference on best linear prediction parameters (ordinary least squares) when the outcome variable is interval valued and the regressors are perfectly measured. [18] extend the familiar Sargan test for overidentifying restrictions to the setting studied by [15]. [25] extend the applicability of [15]'s approach, to cover best linear approximation of any function $f(x)$ that is known to lie within two identified bounding functions. The lower and upper functions defining the band are allowed to be any functions, including ones carrying an index, and can be estimated parametrically or non-parametrically via series methods. [55] proposes an estimator for weighted average derivatives of conditional mean and conditional quantile functionals when either the outcome variable or a regressor is interval-valued.

In contrast, our approach leverages the theory of random sets to propose a local linear interval valued regression estimator for conditional set-valued expectations, and to establish its asymptotic properties. This estimator is novel in the literature, and so are our results establishing its consistency.

The method that we propose also differs significantly from other approaches in the statistical literature; see [83] for a discussion bridging this literature with partial identification. In particular, our proposal is distinct from the large and closely related literature that posits parametric models for set-valued data. In these models tools from interval arithmetic are used to build analogs of the classic linear regression model for perfectly measured data, e.g. by assuming that

$\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{x}_i] = A\boldsymbol{x}_i + B$, where A and B are intervals. See e.g. [35], [44], [45], and [85] among others for a discussion of least squares analysis of this and related models. [71] proposes non-parametric smoothing for this model, by applying weighted least squares to the interval data and then using the resulting intercept as the estimator. [31] discuss different interpretations of set-valued data.

In contrast, we leave the conditional set-valued expectation completely unspecified, and non-parametrically estimate all regression functions compatible with the interval valued data.

Structure of the paper. In Section 3.2 we set up our notation and list some definitions and results from random set theory that we use throughout the paper. Then we briefly review local linear regression with singleton data that our method implicitly applies to each tuple $(\boldsymbol{x}, \boldsymbol{y}) : (\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s., and describe our proposed estimator. In Section 3.5 we derive the asymptotic properties of our estimator. In keeping with the tradition in the statistics literature (e.g., [87]) we first treat the case of deterministic design points, and then the case of random design points. In Section 3.6 we report the results of Monte Carlo experiments and of our empirical illustration.

3.2 Random convex sets and their expectation

We begin with listing our notation and some basic facts in convex geometry and random set theory that we use throughout the paper. We refer to [78] for a thorough account of the theory of random sets.

We use boldface capital letters $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ to denote random compact convex sets, normal font capital letters X, Y, Z and A, B, C to denote deterministic compact convex sets, boldface lower case letters $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ to denote random vectors, and normal font lowercase letters x, y, z to denote deterministic vectors. For $x \in \mathbb{R}$, we denote the positive and negative parts of x respectively by

$$x^+ = \max(0, x), \qquad x^- = -\min(0, x).$$

We let $(\Omega, \mathfrak{F}, \mathbf{P})$ denote a non-atomic probability space on which all random vectors and random sets are defined, where Ω is the space of elementary events equipped with σ-algebra $\mathfrak{F}$ and probability measure $\mathbf{P}$. We denote the Euclidean space by $\mathbb{R}^d$, and equip it with the Euclidean norm (which is denoted by $\|\cdot\|_2$). We denote by $\mathcal{K}(\mathbb{R}^d)$ the collection of compact subsets of $\mathbb{R}^d$. We let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ denote the unit sphere in $\mathbb{R}^d$.

A random compact set $\boldsymbol{Y}$ is a map from $(\Omega, \mathfrak{F}, \mathbf{P})$ to $\mathcal{K}(\mathbb{R}^d)$ such that

$$\{\omega : \ \boldsymbol{Y}(\omega) \cap K \neq \emptyset\} \in \mathfrak{F},$$

for each compact set $K \subset \mathbb{R}^d$. If $\|\boldsymbol{Y}\|_H = \sup\{\|\boldsymbol{y}\|_2 : \ \boldsymbol{y} \in \boldsymbol{Y}\}$ is integrable, the set $\boldsymbol{Y}$ is called *integrably bounded*. Random sets $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are said to be independently and identically distributed if

$$\mathbf{P}(\boldsymbol{Y}_1 \cap K_1 \neq \emptyset, \ldots, \boldsymbol{Y}_n \cap K_n \neq \emptyset) = \prod_{i=1}^{n} \mathbf{P}(\boldsymbol{Y}_i \cap K_i \neq \emptyset),$$

for all $K_1, \ldots, K_n \in \mathcal{K}$ and

$$\mathbf{P}(\boldsymbol{Y}_i \cap K \neq \emptyset) = \mathbf{P}(\boldsymbol{Y}_j \cap K \neq \emptyset),$$

for all $i \neq j \in \{1, \ldots, n\}$ and $K \in \mathcal{K}$.

We define the (Minkowski) sum of two compact sets A and B in $\mathbb{R}^d$ elementwise as

$$A + B = \{x + y : \ x \in A, \ y \in B\}.$$

We let $cA = \{cx : \ x \in A\}$ denote the scaling of A by $c \in \mathbb{R}$, while $-A = \{-x : \ x \in A\}$ is the set centrally symmetric to A. Given a compact convex set (a *convex body*) $A \subset \mathbb{R}^d$, the *support function* of A is

$$s(A, v) = \sup_{a \in A} \langle v, a \rangle, \qquad v \in \mathbb{R}^d,$$

where $\langle v, a \rangle$ denotes the scalar product. If A is convex, its support function uniquely indentifies A, because

$$A = \bigcap_{v \in \mathbb{S}^{d-1}} \{a \in \mathbb{R}^d : \langle v, a \rangle \leq s(A, v)\}. \tag{3.2}$$

Because $s(tA, v) = ts(A, v)$ for $t \geq 0$, $s(-A, v) = s(A, -v)$, and

$$s(A + B, v) = s(A, v) + s(B, v),$$

the support function is often restricted to $v \in \mathbb{S}^{d-1}$. The width function of A is defined by

$$w(A, v) = s(A, v) + s(A, -v) = w(A, -v), \qquad v \in \mathbb{S}^{d-1},$$

and it is easy to see that the width function is non negative. If $d = 1$, then A is a closed interval in $\mathbb{R}$, and the unit sphere $\mathbb{S}^{d-1} = \{-1, 1\}$ consists of two points. In this case, the width function is the length of the interval.

A random convex compact set $\boldsymbol{Y}$ is a map from $(\Omega, \mathfrak{F}, \mathbf{P})$ to $\mathcal{K}_C(\mathbb{R}^d)$ satisfying (3.2), with $\mathcal{K}_C(\mathbb{R}^d)$ denoting the collection of compact and convex subsets of $\mathbb{R}^d$. Its measurability is equivalent to the fact that $s(\boldsymbol{Y}, v)$ is a random variable for each $v \in \mathbb{R}^d$. As specified in Assumption J below, we require $\boldsymbol{Y}$ to be integrably bounded. Because $\|\boldsymbol{Y}\|$ equals the supremum of $s(\boldsymbol{Y}, v)$ for v from the unit sphere, the support function is integrable. The expected support function $\mathbb{E}s(\boldsymbol{Y}, v)$ is the support function of the convex body $\mathbb{E}\boldsymbol{Y}$, which in turn is called the *expectation* of $\boldsymbol{Y}$. Note that $\mathbb{E}w(\boldsymbol{Y}, v) = w(\mathbb{E}\boldsymbol{Y}, v)$. This expectation equals the set of $\mathbb{E}\boldsymbol{y}$ for all

random vectors $\boldsymbol{y}$ such that $y \in \boldsymbol{Y}$ a.s.; in this case $\boldsymbol{y}$ is said to be a (measurable) *selection* of $\boldsymbol{Y}$.

Similarly, for given x it is possible to define the *conditional expectation*

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{x} = x] = \left\{\mathbb{E}[\boldsymbol{y}|\boldsymbol{x} = x] : \boldsymbol{y} \in \boldsymbol{Y} \text{ a.s.}\right\}.$$

Also in this case it holds that $s(\mathbb{E}[\boldsymbol{Y}|\boldsymbol{x} = x], v) = \mathbb{E}[s(\boldsymbol{Y}, v)|\boldsymbol{x} = x]$.

In our analysis, the true but unobservable outcome associated with $\boldsymbol{x} \in \mathbb{R}^m$ is a random vector $\boldsymbol{y}$ that almost surely takes values in $\boldsymbol{Y}$, so that $\boldsymbol{y}$ is a measurable selection of $\boldsymbol{Y}$. The pair $(\boldsymbol{x}, \boldsymbol{y})$ is a selection of $\{\boldsymbol{x}\} \times \boldsymbol{Y}$, a random closed set in $\mathbb{R}^m \times \mathbb{R}^d$.

The family of support functions of all nonempty compact convex subsets in $\mathbb{R}^d$ is a subset of the family of continuous functions on the unit sphere $\mathbb{S}^{d-1}$. In particular, the Hausdorff metric between compact convex sets equals the uniform (L_∞) distance between their support functions. For our purposes, it is convenient to endow the family of continuous functions on the union sphere with the L_2-metric and the corresponding norm that is defined by

$$\|s(A, v)\|_2 = \left[\int_{\mathbb{S}^{d-1}} (s(A, v))^2 \, dv\right]^{\frac{1}{2}}$$

on support functions of compact convex sets A. The integration is performed with respect to the uniform measure on $\mathbb{S}^{d-1}$. Then

$$L(A, B) = \|s(A, v) - s(B, v)\|_2 \tag{3.3}$$

defines an L_2 distance on $\mathcal{K}_C(\mathbb{R}^d)$, which we employ to establish consistency of our estimator. This distance differs from the standard Hausdorff distance used in the related literature in partial identification and in the standard laws of large numbers and central limit theorems for Minkowski averages of random sets.

However, under our assumptions the result of [90, Theorem 3] shows that these two metrics define the same topology, and so the consistency with respect to the L_2 distance implies consistency with respect to the L_∞ distance. At the same time, use of the L_2 distance is particularly well suited to analyze properties of estimators based on least squares minimization.

3.3 Non-parametric smoothing for a given selection $(\boldsymbol{x}, \boldsymbol{y})$

To simplify the exposition, henceforth we assume that $\boldsymbol{x}$ is a scalar random variable taking values in an interval $I \subset \mathbb{R}$. Our results apply, subject only to modification in notation and convergence rates (as in the point identified case), with vector-valued $\boldsymbol{x}$ provided real-valued bandwidth is replaced by a matrix one.

We first focus on a specific selection $(\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s. Such selection is associated with a single function $m(\cdot) \in \mathcal{M}$, and the estimator for this function can be obtained from the classical approach.

In particular, the local polynomial estimator of order p based on observations $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, is obtained by minimizing the weighted least squares

$$\sum_{i=1}^{n} \left(\boldsymbol{y}_i - \theta_0 - \theta_1(\boldsymbol{x}_i - x_0) - \cdots - \theta_p(\boldsymbol{x}_i - x_0)^p \right)^2 K(\frac{\boldsymbol{x}_i - x_0}{h_n}) \tag{3.4}$$

with respect to $\theta_0, \ldots, \theta_p$. The kernel function $K(\cdot)$ is a non-negative integrable function and the tuning parameter h_n is called the *bandwidth*. It is typically assumed that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. The following conditions on the kernel function are imposed throughout this paper.

Assumption I (Kernel function)**.** *The kernel $K(z)$, $z \in \mathbb{R}$, is a non-negative function bounded above by $K_{\max} < \infty$, with compact support $[-c_K, c_K]$ for some $c_K \in (0, \infty)$, and*

satisfying

$$\int K(z)\,dz = 1, \qquad \int zK(z)\,dz = 0,$$

Denote $\mathrm{Var}_K = \int z^2 K(z)\,dz$.

Solving explicitly the weighted least squares minimization problem one obtains the minimizer $\hat{\theta}$, and the first entry of it, the intercept $\hat{\theta}_0$, is used to estimate $m(x_0)$. Such estimator can be written as

$$\hat{m}(x_0) = \sum_{i=1}^{n} \ell_i(x_0)\boldsymbol{y}_i, \tag{3.5}$$

where

$$\begin{aligned}
\ell_i(x_0) &= \frac{1}{nh_n} u^T(0)\mathcal{B}_{nx_0}^{-1} u\left(\frac{\boldsymbol{x}_i - x_0}{h_n}\right) K(\frac{\boldsymbol{x}_i - x_0}{h_n}),\\
u(z) &= \left(1, z, z^2/2!, \ldots, z^p/p!\right)^\top,\\
\mathcal{B}_{nx_0} &= \frac{1}{nh_n}\sum_{i=1}^{n} u\left(\frac{\boldsymbol{x}_i - x_0}{h_n}\right) u^\top\left(\frac{\boldsymbol{x}_i - x_0}{h_n}\right) K\left(\frac{\boldsymbol{x}_i - x_0}{h_n}\right).
\end{aligned}$$

Note that

$$\sum_{i=1}^{n} \ell_i(x_0) = 1. \tag{3.6}$$

Denote

$$\boldsymbol{s}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\kappa}_{in}(\boldsymbol{x}_i - x_0)^j, \qquad j = 0, 1, \ldots,$$

where

$$\boldsymbol{\kappa}_{in} = K(\frac{\boldsymbol{x}_i - x_0}{h_n}).$$

It is easy to see that

$$\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 \geq 0. \tag{3.7}$$

If $p = 0$ (local constant regression), $\hat{M}(x_0)$ is the Nadaraya-Watson estimator

with $\mathcal{B}_{nx_0} = (\boldsymbol{s}_0)$ and $\ell_i(x_0) = \boldsymbol{\kappa}_{in}/(n\boldsymbol{s}_0)$. If $p = 1$ (local linear regression),

$$\mathcal{B}_{nx_0} = \begin{pmatrix} \boldsymbol{s}_0/h_n & \boldsymbol{s}_1/h_n^2 \\ \boldsymbol{s}_1/h_n^2 & \boldsymbol{s}_2/h_n^3 \end{pmatrix},$$

and

$$\ell_i(x_0) = \frac{\boldsymbol{\kappa}_{in}}{n} \frac{\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2}. \tag{3.8}$$

3.4 Non-parametric smoothing for the random set $\{\boldsymbol{x}\} \times \boldsymbol{Y}$

In the following we assume that $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$, is a sample of i.i.d. realizations of $(\boldsymbol{x}, \boldsymbol{Y})$, where $\boldsymbol{Y}$ satisfies Assumption J below. When the outcome data is set-valued, it is necessary to obtain an estimator for the collection of conditional expectations $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x} = x]$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s. This can be accomplished by repeating the procedure in the previous section for all selections of $\{\boldsymbol{x}\} \times \boldsymbol{Y}$. We show that computationally this is easily achieved by taking the following average of the set-valued data:

$$\hat{M}(x_0) = \sum_{i=1}^{n} \ell_i(x_0)\boldsymbol{Y}_i. \tag{3.9}$$

When $p = 0$ we obtain a local constant set-valued regression estimator; when $p = 1$, we obtain a local linear set-valued regression estimator.

The estimator in (3.9) yields a convex set, and therefore in view of (3.2) we can characterize its properties by working with its support function. To simplify notation, in what follows we omit the argument x_0 in $\ell_i(x_0)$ and write shortly ℓ_i, unless the dependence on x_0 is essential. By representing

$$\ell_i = \ell_i^+ - \ell_i^-$$

as the difference of its positive and negative parts, we arrive at

$$\begin{aligned} s(\hat{M}(x_0), v) &= \sum_{i=1}^{n} \ell_i^+ s(\boldsymbol{Y}_i, v) + \sum_{i=1}^{n} \ell_i^- s(\boldsymbol{Y}_i, -v) \\ &= \sum_{i=1}^{n} (\ell_i - \ell_i^-) s(\boldsymbol{Y}_i, v) + \sum_{i=1}^{n} \ell_i^- s(\boldsymbol{Y}_i, -v) \\ &= \sum_{i=1}^{n} \ell_i s(\boldsymbol{Y}_i, v) + \sum_{i=1}^{n} \ell_i^- w(\boldsymbol{Y}_i, v). \end{aligned}$$

Remark 9. *When $d = 1$ and $\boldsymbol{Y} = [\boldsymbol{y}_L, \boldsymbol{y}_U]$ with $\mathbf{P}(\boldsymbol{y}_U \geq \boldsymbol{y}_L) = 1$, one might consider an alternative estimator given by the interval $\hat{N}(x_0) = \left[\sum_{i=1}^n \ell_i y_{iL}, \sum_{i=1}^n \ell_i y_{iU} \right]$. Standard arguments as in [39] yield that this estimator is consistent for*

$$M(x_0) = \mathbb{E}[\boldsymbol{Y} | \boldsymbol{x} = x_0] = \Big[\mathbb{E}[\boldsymbol{y}_L | \boldsymbol{x} = x_0], \mathbb{E}[\boldsymbol{y}_U | \boldsymbol{x} = x_0] \Big].$$

However, this estimator can have large finite sample bias, and even be empty, as illustrated in the following example. Suppose that for i with $\ell_i > 0$ we have $\boldsymbol{y}_{iL} = \boldsymbol{y}_{iU}$ and for i with $\ell_i < 0$ we have $\boldsymbol{y}_{iU} > \boldsymbol{y}_{iL}$. Then

$$\begin{aligned} \sum_{i=1}^{n} \ell_i \boldsymbol{y}_{iL} &= \sum_{i=1}^{n} \ell_i^+ \boldsymbol{y}_{iL} - \sum_{i=1}^{n} \ell_i^- \boldsymbol{y}_{iL} \\ &= \sum_{i=1}^{n} \ell_i^+ \boldsymbol{y}_{iU} - \sum_{i=1}^{n} \ell_i^- \boldsymbol{y}_{iL} \\ &> \sum_{i=1}^{n} \ell_i^+ \boldsymbol{y}_{iU} - \sum_{i=1}^{n} \ell_i^- \boldsymbol{y}_{iU} = \sum_{i=1}^{n} \ell_i \boldsymbol{y}_{iU}. \end{aligned}$$

A similarly empty estimator may result even if $\boldsymbol{y}_{iU} > \boldsymbol{y}_{iL}$ whenever $\ell_i > 0$, depending on the realizations of $\boldsymbol{y}_{iL}$ and $\boldsymbol{y}_{iU}$.

The same effect is the case in dimension $d \geq 2$, where the estimator based on smoothing of support functions in each individual direction may turn out to be empty.

We assume throughout the paper that the observed set-valued responses satisfy the following requirement.

Assumption J (Observed responses)**.** *Conditionally on the design points* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$*, the observations* $\boldsymbol{Y}_i$*,* $i = 1, \ldots, n$*, are random compact convex sets such that*

(i) $s(\boldsymbol{Y}_i, v) = s(M(\boldsymbol{x}_i), v) + \varepsilon_i(v)$*,* $v \in \mathbb{S}^{d-1}$*, where* $\varepsilon_i(\cdot)$*,* $i = 1, \ldots, n$*, are i.i.d. copies of a centered square integrable random function* $\varepsilon(v)$*,* $v \in \mathbb{S}^{d-1}$*, that is independent of the design points and is such that* $\boldsymbol{Y}_i = \bigcap_{v \in \mathbb{S}^{d-1}} \{y : \ y \leq s(\boldsymbol{Y}_i, v)\} \neq \emptyset$ *a.s.*

(ii) $\boldsymbol{Y}_i \subset \xi_i + B$ *a.s. for integrable random vectors* ξ_i*,* $i = 1, \ldots, n$*, and a deterministic compact set* B *that is the same for all* i*.*

Part (i) of Assumption J guarantees that $\boldsymbol{Y}_i$ is almost surely non-empty. It is easiest to interpret in the case $d = 1$. Then the assumption states that $\boldsymbol{y}_{Li} = \mathbb{E}[\boldsymbol{y}_L|\boldsymbol{x}] + \varepsilon_i(-1)$, $\boldsymbol{y}_{Ui} = \mathbb{E}[\boldsymbol{y}_U|\boldsymbol{x}] + \varepsilon_i(1)$, and that $\varepsilon_i(1) - \varepsilon_i(-1) \geq -(\mathbb{E}[\boldsymbol{y}_U|\boldsymbol{x}] - \mathbb{E}[\boldsymbol{y}_L|\boldsymbol{x}])$ a.s. The latter condition replicates the requirement that $\mathbf{P}(\boldsymbol{y}_U \geq \boldsymbol{y}_L) = 1$. Note that ε does not admit a geometric interpretation as the support function of a random set. Part (ii) of Assumption J guarantees that $\boldsymbol{Y}_i$ is uniformly integrably bounded.

Denote

$$C(v, u) = \mathbb{E}[\varepsilon(v)\varepsilon(u)]$$

the covariance function of ε and by $\sigma^2_{\max}$ the maximum of $\mathbb{E}(\varepsilon(v)^2)$ over $v \in \mathbb{S}^{d-1}$. If $\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{x}_i] = M(\boldsymbol{x}_i)$, then

$$\mathbb{E}[s(\hat{M}(x_0), v)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = \sum_{i=1}^{n} \ell_i(s(M(\boldsymbol{x}_i), v) + \sum_{i=1}^{n} \ell_i^- w(M(\boldsymbol{x}_i), v).$$

The quality of the set-valued estimator $\hat{M}(x_0)$ is measured by the quadratic loss function defined in (3.3),

$$L(\hat{M}(x_0), M(x_0))^2 = \int_{\mathbb{S}^{d-1}} (s(\hat{M}(x_0), v) - s(M(x_0), v))^2 \, dv.$$

The mean squared error (MSE) of the estimator is then the expectation of $L(\hat{M}(x_0), M(x_0))$. A classic bias and variance decomposition yields

$$\mathrm{MSE}(x_0) = \int_{\mathbb{S}^{d-1}} b^2_{x_0}(v)\, dv + \int_{\mathbb{S}^{d-1}} \sigma^2_{x_0}(v)\, dv,$$

where $b^2_{x_0}(v)$ and $\sigma^2_{x_0}(v)$ are squared bias and variance given by

$$b^2_{x_0}(v) = \mathbb{E}\left[\mathbb{E}[s(\hat{M}(x_0), v)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]) - s(M(x_0), v)\right]^2,$$
$$\sigma^2_{x_0}(v) = \mathbb{E}\left[s(\hat{M}(x_0), v) - s(\mathbb{E}[\hat{M}(x_0)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n], v)\right]^2.$$

Rearranging the terms, we arrive at

$$b^2_{x_0}(v) = \mathbb{E}\left(\sum_{i=1}^n \ell_i(s(M(\boldsymbol{x}_i), v) - s(M(x_0), v)) + \sum_{i=1}^n \ell_i^- w(M(\boldsymbol{x}_i), v)\right)^2 \tag{3.10}$$

and

$$\sigma^2_{x_0}(v) = \mathbb{E}\left(\sum_{i=1}^n \ell_i(s(\boldsymbol{Y}_i, v) - s(M(\boldsymbol{x}_i), v)) + \sum_{i=1}^n \ell_i^- (w(\boldsymbol{Y}_i, v) - w(M(\boldsymbol{x}_i), v))\right)^2.$$

By Assumption J, the variance can be expressed as

$$\sigma^2_{x_0}(v) = \mathbb{E}\left(\sum_{i=1}^n \ell_i \varepsilon_i(v) + \sum_{i=1}^n \ell_i^- (\varepsilon_i(v) + \varepsilon_i(-v))\right)^2. \tag{3.11}$$

Differently from the classical case with singleton response $\boldsymbol{y}_i$, the negative parts of the weights in (C.1) play an essential role for set-valued response. This because while the difference between $s(M(\boldsymbol{x}_i), v)$ and $s(M(x_0), v)$ is small when $\boldsymbol{x}_i$ is close to x_0, the width $w(M(\boldsymbol{x}_i), v)$ does not vanish as $\boldsymbol{x}_i$ becomes closer to x_0. Thus, the bias increases by a constant and may not tend to zero if some weights are negative and not close to zero.

3.5 Asymptotic properties of the set-valued estimators

In the local linear regression setting, negative weights may appear in (3.10) and hence affect the bias in the case of set-valued data. In Proposition 1 below we

determine the contribution to the bias resulting from the negative weights. In working with random designs, we assume $I = \mathbb{R}$ and impose the following condition.

Assumption K (Theoretical response function)**.** *The function $M(x)$, $x \in \mathbb{R}$, is such that $s(M(x), v)$ admits the second derivative $s''(M(x), v)$ in x which is uniformly bounded for all $v \in \mathbb{S}^{d-1}$.*

Furthermore, we assume that the common density f of the independent design points satisfies the following condition which is similar to that imposed in [39, Condition 1(ii)] with singleton-valued responses.

Assumption L (Density)**.** *The density f is strictly positive at x_0 and belongs to the family $\mathcal{H}(1, \gamma)$ of Lipschitz functions with constant $\gamma > 0$, that is*

$$|f(x') - f(x'')| \leq \gamma |x' - x''|$$

for all $x', x'' \in \mathbb{R}$.

Following [39], in order to avoid zero in the denominator of the local linear estimator, we redefine ℓ_i by letting

$$\ell_i = \frac{\boldsymbol{\kappa}_{in}}{n} \frac{\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1}{\boldsymbol{s}_2 \boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}}.$$

For a sequence $\{z_n, n \geq 1\}$ of random variables determined through the design points and the observations, write $z_n = O_r(a_n)$ if

$$\sup_{f \in \mathcal{H}(1,\gamma)} \mathbb{E}|z_n|^r = O(a_n^r),$$

see [39]. The notation $o_r(a_n)$ is defined similarly. As stated in [39], $O_r(a_n)O_r(b_n) = O_{r/2}(a_n b_n)$, and

$$z_n = \mathbb{E}z_n + O_r(\mathbb{E}|z_n - \mathbb{E}z_n|^r)^{1/r}. \tag{3.12}$$

We use o and O to denote the deterministic order of magnitude uniformly in $f \in \mathcal{H}(1, \gamma)$.

Proposition 1. *Under Assumptions I and L, for some $r \geq 1$,*

$$\mathbb{E}\left(\sum_{i=1}^{n} \ell_i^-\right)^2 = o(\tilde{h}_n^r) \quad as \;\; n \to \infty,\; h_n \to 0,\; nh_n \to \infty,$$

where

$$\tilde{h}_n = h_n + \frac{1}{\sqrt{nh_n}}, \quad n \geq 1. \tag{3.13}$$

Proof. Since the kernel is assumed to have a compact support, the moments $\int z^{2r} K(z) dz$ exist for all $r > 0$. According to [39, Eq. (6.4)], for an integer $r \geq 1$,

$$\boldsymbol{s}_j = \mathbb{E}\boldsymbol{s}_j + h_n^{j+1} O_r(\frac{1}{\sqrt{nh_n}}), \quad j = 0, 1, 2, \tag{3.14}$$

as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$. The expectations $\mathbb{E}\boldsymbol{s}_j$ can be calculated as follows:

$$\begin{aligned}
\mathbb{E}\boldsymbol{s}_0 &= h_n \int K(z) f(zh_n + x_0)\, dz = h_n \int K(z)(f(x_0) + O(h_n))\, dz \\
&= h_n[f(x_0) + O(h_n)], \\
\mathbb{E}\boldsymbol{s}_1 &= h_n^2 \int zK(z) f(zh_n + x_0)\, dz = h_n^2 \int zK(z)(f(x_0) + O(h_n))\, dz \\
&= h_n^2 O(h_n), \\
\mathbb{E}\boldsymbol{s}_2 &= h_n^3 \int z^2 K(z) f(zh_n + x_0)\, dz = h_n^3 \int z^2 K(z)(f(x_0) + O(h_n))\, dz \\
&= h_n^3 (f(x_0) \operatorname{Var}_K + O(h_n)).
\end{aligned}$$

In view of (3.14), for an integer $r \geq 1$,

$$\boldsymbol{s}_j = h_n^{j+1} \left(f(x_0) \int z^j K(z)\, dz + O_r(\tilde{h}_n) \right), \quad j = 0, 1, 2. \tag{3.15}$$

Thus

$$\boldsymbol{s}_0 = h_n f(x_0)(1 + o_r(1)), \tag{3.16}$$

$$\boldsymbol{s}_1 = h_n^2 o_r(1) \tag{3.17}$$

$$\boldsymbol{s}_2 = h_n^3 f(x_0) \operatorname{Var}_K (1 + o_r(1)) \tag{3.18}$$

If $|\boldsymbol{x}_i - x_0| \le c_K h_n$, then, uniformly in i,

$$\begin{aligned}
\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1 &= h_n^3 f(x_0) \operatorname{Var}_K + h_n^3 O_r(\tilde{h}_n) - (\boldsymbol{x}_i - x_0) h_n^2 O_r(\tilde{h}_n) \\
&= h_n^3 f(x_0) \operatorname{Var}_K + h_n^3 O_r(\tilde{h}_n) + O(h_n) h_n^2 O_r(\tilde{h}_n) \\
&= h_n^3 f(x_0) \operatorname{Var}_K + h_n^3 O_r(\tilde{h}_n).
\end{aligned}$$

Thus, $\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1 < 0$ implies that the event

$$D_n = \{ f(x_0) \operatorname{Var}_K < O_r(\tilde{h}_n) \}$$

occurs. Since D_n is independent of i, the Hölder inequality yields that

$$\mathbb{E}\left(\sum_{i=1}^n \ell_i^- \right)^2 = \mathbb{E}\left(- \sum_{i=1}^n \ell_i \mathbf{1}_{D_n} \right)^2 \le \left(\mathbb{E}\left(\sum_{i=1}^n \ell_i \right)^4 \mathbf{P}(D_n) \right)^{1/2}. \tag{3.19}$$

The Chebyshev inequality and the definition of $O_r(a_n)$ yield that

$$\mathbf{P}(D_n) \le \frac{1}{f(x_0)^r \operatorname{Var}_K^r} \mathbb{E} \left| O_r(\tilde{h}_n) \right|^r \le \frac{c_r \tilde{h}_n^r}{f(x_0)^r \operatorname{Var}_K^r}$$

for a positive constant c_r. By repeatedly choosing sufficiently large values of r (that may vary across our uses of it), we obtain

$$\sqrt{\mathbf{P}(D_n)} = o(\tilde{h}_n^r). \tag{3.20}$$

Then

$$\sum_{i=1}^n \ell_i = \frac{\sum_{i=1}^n \boldsymbol{\kappa}_{in} \boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1}{\boldsymbol{s}_2 \boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}} \tag{3.21}$$

$$= \frac{\boldsymbol{s}_2 \boldsymbol{s}_0 - \boldsymbol{s}_1^2}{\boldsymbol{s}_2 \boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}} \tag{3.22}$$

Following the arguments used to derive [39, Eq. (6.6)],

$$\frac{h_n^4}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}} = \frac{1}{f(x_0)^2 \operatorname{Var}_K} + o_r(1), \tag{3.23}$$

for sufficiently large r. In view of (3.16), (3.17), and (3.18),

$$\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 = h_n^4 f(x_0) \operatorname{Var}_K(1 + o_r(1)).$$

Therefore,

$$\sum_{i=1}^{n} \ell_i = 1 + o_r(1).$$

Choosing $r \geq 4$ yields that

$$\mathbb{E}\left(\sum_{i=1}^{n} \ell_i^-\right)^4 = \mathbb{E}[1 + o_r(1)]^4 = 1 + o(1). \tag{3.24}$$

Substituting (3.20) and (3.24) into (3.19), we have

$$\mathbb{E}\left(\sum_{i=1}^{n} \ell_i^-\right)^2 \leq (1 + o(1))\, o(\tilde{h}_n^r) = o(\tilde{h}_n^r). \qquad \square$$

Theorem 10. *Under Assumptions I, J, K, and L, if $h_n = \alpha n^{-\beta}$ with $0 < \beta < 1$, the mean squared error of the local linear estimator in* (3.9) *with $p = 1$ is*

$$\operatorname{MSE}(x_0) = \frac{h_n^4 (\operatorname{Var}_K)^2}{4} \int_{\mathbb{S}^{d-1}} s''(M(x_0), v)^2 \, dv + \frac{1}{n h_n f(x_0)} \int_{\mathbb{S}^{d-1}} C(v, v) \, dv \int K^2(z) \, dz + o(h_n^4 + \frac{1}{n h_n}).$$

Proof. The squared bias can be written as

$$b_{x_0}^2(v) = \mathbb{E}[(B_1 + B_2)^2], \tag{3.25}$$

where

$$B_1 = \sum_{i=1}^{n} \ell_i (m_v(\boldsymbol{x}_i) - m_v(x_0)),$$

$$B_2 = \sum_{i=1}^{n} \ell_i^- w(M(\boldsymbol{x}_i), v).$$

According to [39, p. 212],

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\kappa}_{in}(\boldsymbol{s}_2-(\boldsymbol{x}_i-x_0)\boldsymbol{s}_1)(m_v(\boldsymbol{x}_i)-m_v(x_0))\\
&\qquad=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\kappa}_{in}(\boldsymbol{s}_2-(\boldsymbol{x}_i-x_0)\boldsymbol{s}_1)(m_v(\boldsymbol{x}_i)-m_v(x_0))+s'(M(x_0),v)(\boldsymbol{x}_i-x_0))\\
&\qquad=h_n^6 f(x_0)\operatorname{Var}_K a_n+o_4(h_n^6),
\end{aligned}$$

where

$$a_n=h_n^{-3}\mathbb{E}\left(m_v(\boldsymbol{x})-m_v(x_0)-s'(M(x_0),v)(\boldsymbol{x}-x_0)K(\frac{\boldsymbol{x}-x_0}{h_n})\right).$$

By (3.23), and using the definition of o_r, we have

$$\begin{aligned}
\mathbb{E}B_1^2&=\mathbb{E}\left(\frac{\sum_{i=1}^{n}\boldsymbol{\kappa}_{in}(\boldsymbol{s}_2-(\boldsymbol{x}_i-x_0)\boldsymbol{s}_1)(m_v(\boldsymbol{x}_i)-m_v(x_0))}{\boldsymbol{s}_2\boldsymbol{s}_0-\boldsymbol{s}_1^2+n^{-4}}\right)^2\\
&=\left(\frac{U_n}{f(x_0)}\right)^2 h_n^4+o(h_n^4),
\end{aligned}\tag{3.26}$$

where

$$U_n=h_n^{-2}\left(\frac{1}{2}s''(M(x_0),v)\operatorname{Var}_K f(x_0)h_n^2+o(h_n^2)\right)\tag{3.27}$$

by the Taylor expansion. Substituting (3.27) into (3.26) yields that

$$\mathbb{E}B_1^2=\frac{1}{4}s''(M(x_0),v)^2(\operatorname{Var}_K)^2h_n^4+o(h_n^4).\tag{3.28}$$

A more detailed calculation could be found in the proof of [39, Theorem 3]. Furthermore,

$$\mathbb{E}B_2^2\le w_{\max}^2\mathbb{E}\left(\sum_{i=1}^{n}\ell_i^-\right)^2=o(\tilde{h}_n^{10}),\tag{3.29}$$

where $w_{\max}$ is a finite deterministic bound on the width of $M(\boldsymbol{x})$ in any direction $v\in\mathbb{S}^{d-1}$ resulting from Assumption J.

Using Cauchy-Schwarz inequality, (3.29) and (3.28),

$$\begin{aligned}
\mathbb{E}(B_1B_2)&\le\sqrt{\mathbb{E}B_1^2\mathbb{E}B_2^2}=\left(\frac{1}{4}s''(M(x_0),v)^2(\operatorname{Var}_K)^2h_n^4+o(h_n^4)\right)^{1/2}\ o(\tilde{h}_n^5)\\
&=O(h_n^2)o(h_n^4+\frac{1}{nh_n})=o(h_n^4+\frac{1}{nh_n}).
\end{aligned}\tag{3.30}$$

Here we have used that

$$o(\tilde{h}_n^5) = o(h_n + \frac{1}{nh_n})$$

with the choice $h_n = \alpha n^{-\beta}$. Substituting equations (3.28),(3.30) and (3.29) into (3.25), we have

$$\int_{\mathbb{S}^{d-1}} b_{x_0}^2(v)\, dv = \frac{1}{4} \int_{\mathbb{S}^{d-1}} s''(M(x_0), v)^2\, dv (\mathrm{Var}_K)^2 h_n^4 + o(h_n^4 + \frac{1}{nh_n}). \tag{3.31}$$

Now we bound the variance of the estimator splitting (3.11) into the sum of three terms. The first term is

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^n \ell_i \varepsilon_i(v)\right)^2 &= \mathbb{E}\left(C(v,v)\sum_{i=1}^n \ell_i^2\right) \\
&= \mathbb{E}\left(\frac{C(v,v)\sum_{i=1}^n \boldsymbol{\kappa}_{in}^2(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)^2}{n^2(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2}\right) \\
&= \mathbb{E}\left(\frac{C(v,v)(\boldsymbol{s}_2^2\boldsymbol{s}_0^* - 2\boldsymbol{s}_2\boldsymbol{s}_1\boldsymbol{s}_1^* + \boldsymbol{s}_1^2\boldsymbol{s}_2^*)}{n(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2}\right) \\
&= \mathbb{E}\left(\frac{C(v,v)\boldsymbol{s}_2^2\boldsymbol{s}_0^*}{n(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2}\right) + \mathbb{E}\left(\frac{C(v,v)(-2\boldsymbol{s}_2\boldsymbol{s}_1\boldsymbol{s}_1^* + \boldsymbol{s}_1^2\boldsymbol{s}_2^*)}{n(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2}\right),
\end{aligned} \tag{3.32}$$

where

$$\boldsymbol{s}_j^* = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\kappa}_{in}^2 (x_i - x_0)^j, \quad j = 0, 1, 2.$$

Following [39, Eq. (6.13)],

$$\boldsymbol{s}_j^* = h_n^{j+1}\left(f(x_0)\int z^j K^2(z)\, dz + o_4(1)\right).$$

Furthermore, (3.15) implies that

$$\sigma^2\boldsymbol{s}_2^2\boldsymbol{s}_0^* = h_n^7 f^3(x_0)(\mathrm{Var}_K)^2\sigma^2 \int K^2(z)\, dz + h_n^7 o_2(1).$$

Combining this with (3.23),

$$\begin{aligned}
\mathbb{E}\left(\frac{\sigma^2\boldsymbol{s}_2^2\boldsymbol{s}_0^*}{n(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2}\right) &= \frac{h_n^7 f^3(x_0)(\mathrm{Var}_K)^2 C(v,v)\int K^2(z)\, dz}{nh_n^8 f^4(x_0)(\mathrm{Var}_K)^2} + \frac{h_n^7}{nh_n^8}o(1) \\
&= \frac{C(v,v)\int K^2(z)\, dz}{nh_n f(x_0)} + o(\frac{1}{nh_n}).
\end{aligned}$$

Since $\int zK(z)\,dz = 0$,

$$-2\boldsymbol{s}_2\boldsymbol{s}_1\boldsymbol{s}_1^* = h_n^7(f(x_0)\operatorname{Var}_K + o_8(1))o_8(1)(f(x_0)\int z^j K^2(z)\,dz + o_4(1))$$
$$= h_n^7 o_2(1).$$

Analogously,

$$\boldsymbol{s}_1^2\boldsymbol{s}_2^* = h_n^7 o_2(1).$$

Both these terms are as small as the minor term of $\boldsymbol{s}_2^2\boldsymbol{s}_0^*$. Therefore, (3.32) is dominated by its first term.

$$\mathbb{E}\left(\sum_{i=1}^n \ell_i\varepsilon_i(v)\right)^2 = \frac{C(v,v)\int K^2(z)\,dz}{nh_n f(x_0)} + o(\frac{1}{nh_n}) \tag{3.33}$$

The second term is

$$\mathbb{E}\sum_{1\le i<j\le n} \ell_i\ell_j^-\varepsilon_i(v)(\varepsilon_j(v) + \varepsilon_j(-v)) = 0.$$

Finally, consider

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^n \ell_i^-(\varepsilon_i(v) + \varepsilon_i(-v))\right)^2 &= (C(v,v) + 2C(v,-v) + C(-v,-v))\mathbb{E}\sum_{i=1}^n(\ell_i^-)^2\\ &\le 4\sigma_{\max}^2\mathbb{E}\sum_{i=1}^n(\ell_i^-)^2 \le 4\sigma_{\max}^2\mathbb{E}\left(\sum_{i=1}^n \ell_i^-\right)^2\\ &= 4\sigma_{\max}^2 o(\tilde{h}_n^{10}) = o(h_n^4 + \frac{1}{nh_n}).\end{aligned}$$

Therefore,

$$\int_{\mathbb{S}^{d-1}} \sigma_{x_0}^2(v)\,dv = \frac{1}{nh_n f(x_0)}\int_{\mathbb{S}^{d-1}} C(v,v)\,dv\int K^2(z)\,dz + o(h_n^4 + \frac{1}{nh_n}),$$

and the result follows by adding (3.31) to it. □

The following result establishes a limit theorem for the support function of the estimators as processes on the unit sphere. Let $\xi(v)$, $v \in \mathbb{S}^{d-1}$, be a centered Gaussian process on the unit sphere with the covariance

$$\mathbb{E}(\xi(v)\xi(u)) = \frac{C(v,u)}{f(x_0)}\int K(z)^2\,dz.$$

Theorem 11. *Assume that $h_n = \alpha n^{-\beta}$ with $0 < \beta < 1$ and fix $x_0 \in I$. Under Assumptions I, J, K, and L, the stochastic process*

$$\sqrt{nh_n}\left(s(\hat{M}(x_0), v) - s(M(x_0), v) - h_n^2 \frac{1}{2} s''(M(x_0), v)\sigma_K^2\right)$$

constructed using local linear estimator (3.9) *converges in distribution in the space of continuous functions on $\mathbb{S}^{d-1}$ with the uniform metric to the Gaussian process ξ.*

Proof. It suffices to establish the convergence of one-dimensional distributions; the weak convergence of finite dimensional distributions follows from the Cramér–Wold device, and the functional convergence is established by bounding the Lipschitz constants of the processes as in in [78, Theorem 3.2.1].

First, decompose

$$\begin{aligned} s(\hat{M}, v) - s(M(x_0), v) &= \sum_{i=1}^{n} \ell_i s(\boldsymbol{Y}_i, v) + \sum_{i=1}^{n} \ell_i^- w(\boldsymbol{Y}_i, v) - s(M(x_0), v) \\ &= \sum_{i=1}^{n} \ell_i s(M(\boldsymbol{x}_i), v) + \sum_{i=1}^{n} \ell_i \varepsilon_i(v) + \sum_{i=1}^{n} \ell_i^- w(\boldsymbol{Y}_i, v) - s(M(x_0), v) \end{aligned} \tag{3.34}$$

By Proposition 1 and the fact that L_1-convergence implies the convergence in probability, First, we are going to show that $\sum_{i=1}^{n} \ell_i^- w(\boldsymbol{Y}_i, v) = o_p\left(\frac{1}{\sqrt{nh_n}}\right)$.

$$\sum_{i=1}^{n} \ell_i^- w(\boldsymbol{Y}_i, v) \le w_{\max} \sum_{i=1}^{n} \ell_i^- = w_{\max} o_p\left(\tilde{h}_n^r\right),$$

where $\tilde{h}_n$ is given by (3.13). By choosing r large enough,

$$\sum_{i=1}^{n} \ell_i^- w(\boldsymbol{Y}_i, v) = o_p\left(\tilde{h}_n^r\right) = o_p\left(\frac{1}{\sqrt{nh_n}}\right) \tag{3.35}$$

Using Taylor expansion,

$$s(M(\boldsymbol{x}_i), v) = s(M(x_0), v) + (\boldsymbol{x}_i - x_0)s'(M(x_0), v) + \frac{1}{2}(\boldsymbol{x}_i - x_0)^2 s''(M(x_0), v) + R(x_0, \boldsymbol{x}_i, v),$$

where the rest term $R(x_0, \boldsymbol{x}_i, v)$ is of a smaller oder than $\frac{1}{2}(\boldsymbol{x}_i - x_0)^2 s''(M(x_0), v)$. Since the local linear estimator satisfies $\sum_{i=1}^n \ell_i(\boldsymbol{x}_i - x_0) = 0$, we have

$$\begin{aligned}
&\sum_{i=1}^n \ell_i s(M(\boldsymbol{x}_i), v) + \sum_{i=1}^n \ell_i \varepsilon_i(v) - s(M(x_0), v) \\
&= \sum_{i=1}^n \ell_i (s(M(\boldsymbol{x}_i), v) - s(M(x_0), v)) - \frac{n^{-4}}{S_2 S_0 - S_1^2 + n^{-4}} s(M(x_0), v) + \sum_{i=1}^n \ell_i \varepsilon_i(v) \\
&= \sum_{i=1}^n \ell_i \left(\frac{1}{2}(\boldsymbol{x}_i - x_0)^2 s''(M(x_0), v) + R(x_0, \boldsymbol{x}_i, v) + \varepsilon_i(v) \right) - \frac{n^{-4}}{S_2 S_0 - S_1^2 + n^{-4}} s(M(x_0), v).
\end{aligned}$$

Since

$$Z_n = \mathbb{E}(Z_n) + O_p(\sqrt{\mathrm{Var}(Z_n)})$$

for a sequence of $\{Z_n, n \geq 1\}$ of random variables with finite variance, it is not difficult to show

$$S_j = h_n^{j+1} f(x_0) \int z^j f(z)\, dz (1 + o_p(1)), \text{ for } j = 0, 1, 2, 3. \tag{3.36}$$

Along with the fact that

$$S_2 S_0 - S_1^2 + n^{-4} = h_n^4 \,\mathrm{Var}_K\, f^2(x_0)(1 + o_p(1)), \tag{3.37}$$

we have

$$\frac{n^{-4}}{S_2 S_0 - S_1^2 + n^{-4}} s(M(x_0), v) = O_p\left(\frac{1}{n^4 h_n^4}\right) = o_p\left(\frac{1}{n^3 h_n^3}\right). \tag{3.38}$$

Combining (3.36) and (3.37), we have

$$\begin{aligned}
&\sum_{i=1}^n \ell_i \left(\frac{1}{2}(\boldsymbol{x}_i - x_0)^2 s''(M(x_0), v) + R(x_0, \boldsymbol{x}_i, v) + \varepsilon_i(v) \right) \\
&= \frac{1}{2}(S_2^2 - S_3 S_1) s''(M(x_0), v) + \frac{1}{n} \sum_{i=1}^n \kappa_i (S_2 - (\boldsymbol{x}_i - x_0) S_1) \varepsilon_i(v) (S_2 S_0 - S_1^2 + n^{-4})^{-1} \\
&= \frac{1}{2} \,\mathrm{Var}_K\, s''(M(x_0), v) h_n^2 (1 + o_p(1)) + \frac{1}{n h_n f(x_0)} \sum_{i=1}^n \kappa_i \varepsilon_i(v)(1 + o_p(1))
\end{aligned} \tag{3.39}$$

By the central limit theorem,

$$\frac{1}{\sqrt{n h_n}} \sum_{i=1}^n \kappa_i \varepsilon_i \tag{3.40}$$

converges in distribution to the centered normal random variable with variance equal to that of $\xi(v)$. The combination of (3.34), (3.35), (3.37), (3.39) and (3.40) yields the result. □

3.6 Monte Carlo Experiments and Empirical Illustration

3.6.1 Cross validation

In the classic setting where the observation pairs (x_i, y_i) are real valued, we generally choose the bandwidth h so that it minimize the leave-one-out cross-validation score which is defined as follows

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{(-i)}(x_i))^2,$$

where

$$\hat{m}_{(-i)}(x) = \sum_{j=1}^{n} y_j \ell_{j,(-i)}(x)$$

$$\ell_{j,(-i)}(x) = \begin{cases} 0 & \text{if } j = i \\ \frac{\ell_j(x)}{\sum_{k\neq i}\ell_k(x)} & \text{if } j \neq i \end{cases}$$

In other words we set the weight on x_i to 0 and renormalize the other weights to sum to one.

Following the same idea, we define the cross-validation score for the set-valued observation pairs $(\boldsymbol{x}_i, \boldsymbol{Y}_i) \in \mathbb{R} \times \mathcal{K}(\mathbb{R}^d)$ as

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\int_{v\in\mathbb{S}^{d-1}} (s(\boldsymbol{Y}_i, v) - s(\hat{M}_{(-i)}(\boldsymbol{x}_i), v))^2 \, d\mu(v), \tag{3.41}$$

where

$$\hat{M}_{(-i)}(x) = \sum_{j=1}^{n} \boldsymbol{Y}_j \ell_{j,(-i)}(x).$$

In case of intervals $\boldsymbol{Y}_i = [\boldsymbol{y}_{iL}, \boldsymbol{y}_{iU}]$, (3.41) turns into

$$\text{CV} = \frac{1}{n} \sum_{i=1}^{n} \frac{(\boldsymbol{y}_{Li} - \hat{M}_{(-iL)}(\boldsymbol{x}_i))^2 + (\boldsymbol{y}_{iU} - \hat{M}_{(-iU)}(\boldsymbol{x}_i))^2}{2},$$

where $\hat{M}_{(-iL)}(\boldsymbol{x}_i)$ and $\hat{M}_{(-iU)}(\boldsymbol{x}_i)$ denote the lower and upper bounds of $\hat{M}_{(-i)}(\boldsymbol{x}_i)$.

APPENDIX A

CHAPTER 1 OF APPENDIX

A.1 Proofs

Lemma 2. *Write* $D_n = (I - M_n \circ \eta)^{-} X_n\beta + (I - M_n \circ \eta)^{-}\epsilon = f(M_n \circ X_n) + \epsilon_1$. *In the first stage problem (19), under assumption 5 and square root lasso regularization parameter* $\lambda \geq c\Lambda/n$,

$$\frac{\|\hat{D}_n - f\|_2}{\sqrt{n}} \leq Cs^{1/2}\lambda$$

where,

$$\Lambda = n\left\|\nabla\sqrt{\hat{Q}(\beta^0)}\right\|_\infty = \max_{1\leq i\leq p}\left\{\frac{\sqrt{n}((X^i)'\epsilon)}{\|\epsilon\|_2}\right\}$$

Lemma one is the same as Theorem 1 in [13]

A.1.1 Theorem 1

Proof. In the second stage, $\hat{D}_n$ is used to replace D_n

$$(\hat{\beta}, \hat{\eta}) = \arg\min_{\beta,\eta}(\|D_n - X_n\beta - (M_n \circ \hat{D}_n)\eta\|_2/\sqrt{n} + 2\lambda\|\eta\|_1/\sqrt{n})$$

take the derivative for the second equation:

$$-(M_n \circ \hat{D}_n)'(D_n - X_n\hat{\beta} - (M_n \circ \hat{D}_n)\hat{\eta})/n + \hat{Q}\lambda\hat{\kappa} = 0 \quad (A)$$

$$-X_n'(D_n - X_n\hat{\beta} - (M_n \circ \hat{D}_n)\hat{\eta})/n = 0 \quad (B)$$

where $\hat{Q} = \|D_n - X_n\hat{\beta} - (M_n \circ \hat{D}_n)\hat{\eta}\|_2$

Substitute $D_n = X_n\beta_0 + (M_n \circ D_n)\eta_0 + \epsilon_n$. Equation (A) can be transformed as:

$$\frac{1}{n}(M_n \circ \hat{D}_n)' X_n(\hat{\beta} - \beta_0) + \frac{1}{n}(M_n \circ \hat{D}_n)'\Big((M_n \circ \hat{D}_n)\hat{\eta} - (M_n \circ D_n)\eta_0\Big) + \hat{Q}\lambda\hat{\kappa} = \frac{(M_n \circ \hat{D}_n)'\epsilon}{n}$$

and further that:

$$\frac{1}{n}(M_n \circ \hat{D}_n)' X_n(\hat{\beta} - \beta_0) + \frac{1}{n}(M_n \circ \hat{D}_n)'(M_n \circ \hat{D}_n)(\hat{\eta} - \eta_0) + \underbrace{\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (\hat{D}_n - D_n)\Big)\eta_0}_{(C)} + Q\lambda\hat{\kappa} = \frac{(M_n \circ \hat{D}_n)'\epsilon}{n}$$

Equation (C) can be written as:

$$\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (\hat{D}_n - D_n)\Big)\eta_0 = \frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (\hat{D}_n - f)\Big)\eta_0 + \frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (f - D_n)\Big)\eta_0$$

And

$$\begin{aligned}\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (f - D_n)\Big)\eta_0 &= -\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ \epsilon_1\Big)\eta_0 \\ &= -\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ \eta^0\Big)(I - M_n \circ \eta_0)^{-1}\epsilon \\ &= -\frac{1}{n}(M_n \circ \hat{D}_n)'\left((I - M_n \circ \eta_0)^{-1} - I\right)\epsilon\end{aligned}$$

Thus (A) is equivlent to:

$$\frac{1}{n}(M_n \circ \hat{D}_n)' X_n(\hat{\beta} - \beta_0) + \frac{1}{n}(M_n \circ \hat{D}_n)'(M_n \circ \hat{D}_n)(\hat{\eta} - \eta_0) + \frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (\hat{D}_n - f)\Big)\eta_0 + \hat{Q}\lambda\hat{\kappa} = \frac{1}{n}(M_n \circ \hat{D}_n)'(I - M_n \circ \eta_0)^{-1}\epsilon$$

Notice that:

$$\begin{aligned}\left\|\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (\hat{D}_n - f)\Big)\eta_0\right\|_\infty &\leq \left\|\frac{1}{n}M_n'(M_n \circ \eta_0)(\hat{D}_n - f)\right\|_\infty \|\hat{D}_n\|_\infty \\ &\leq \frac{1}{n}\left\|M_n'(M_n \circ \eta_0)\right\|_{op2}\left\|(\hat{D}_n - f)\right\|_2 \|\hat{D}_n\|_\infty\end{aligned}$$

where $\|.\|_{op2}$ is the operation norm of the matrix in $l_2 \to l_\infty$ space, which is the maximum l_2 norm of the row.

From lemma 1,

$$\|(\hat{D}_n - f)\|_2 = O\left(\sqrt{s \log n}\right) = o(n^{1/4})$$

Since each entry of M_n is either 1 or 0, and η_0 has $o(\frac{\sqrt{n}}{\log n})$ non-zero entries. $\|M_n \circ \eta_0\|_{op2} \leq o(\frac{n^{1/4}}{\sqrt{\log n}})$. By assumption 5, $\|M_n\|_{op2} = O(\sqrt{\log n})$

$$\left\|M_n'(M_n \circ \eta_0)\right\|_{op2} \leq \|M_n\|_{op2} \|M_n \circ \eta_0\|_{op2} = o(n^{1/4})$$

And $\|\eta_0\|_\infty < 1$:

$$\left\|\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(M_n \circ (\hat{D}_n - f)\Big)\eta_0\right\|_\infty = o(1/\sqrt{n})$$

Similarly (B) can be transformed in the same way, so (A) and (B) are:

$$\begin{aligned}
&\frac{1}{n}(M_n \circ \hat{D}_n)' X_n(\hat{\beta} - \beta_0) + \frac{1}{n}(M_n \circ \hat{D}_n)'(M_n \circ \hat{D}_n)(\hat{\eta} - \eta_0) \\
&\qquad + \hat{Q}\lambda\hat{\kappa} + o(1/\sqrt{n}) = \frac{(M_n \circ \hat{D}_n)'\epsilon_1}{n} \quad (A') \\
&\frac{1}{n}X_n' X_n(\hat{\beta} - \beta_0) + \frac{1}{n}X_n'(M_n \circ \hat{D}_n)(\hat{\eta} - \eta_0) + o(1/\sqrt{n}) \\
&\qquad = \frac{X_n'\epsilon_1}{n} \quad (B')
\end{aligned}$$

From (B′)

$$(\hat{\beta} - \beta_0) = (X_n'X_n)^- X_n'\epsilon_1 - (X_n'X_n)^- X_n'(M_n \circ \hat{D}_n)(\hat{\eta} - \eta_0)$$

And substitute this into (A′)

$$\begin{aligned}
&\frac{1}{n}(M_n \circ \hat{D}_n)'\Big(I - X_n(X_n'X_n)^- X_n'\Big)(M_n \circ \hat{D}_n)(\hat{\eta} - \eta_0) + \hat{Q}\lambda\hat{\kappa} \\
&\qquad = \frac{1}{n}(M_n \circ \hat{D}_n)'\Big(I - X_n(X_n'X_n)^- X_n'\Big)\epsilon_1
\end{aligned}$$

Define $W_n = \left(I - X_n(X_n'X_n)^{-}X_n'\right)$,

$$\frac{1}{n}\tilde{X}_1'\tilde{X}_1(\hat{\eta} - \eta_0) + \hat{Q}\lambda\hat{\kappa} = \frac{1}{n}\tilde{X}_1'\epsilon_1$$

where $\tilde{X}_1 = W_n(M_n \circ \hat{D}_n)$.

Define $\hat{\Theta}$ generated from the nodewise regression on $\tilde{X}_1$ as in [77]. $\hat{\Theta}$ is a reason able approximation to the inverse of $\tilde{X}_1'\tilde{X}_1/n$. Thus,

$$\hat{\eta} - \eta_0 + \hat{\Theta}\hat{Q}\lambda\hat{\kappa} = \frac{1}{n}\hat{\Theta}\tilde{X}_1'\epsilon_1 - \Delta/\sqrt{n}$$

where

$$\Delta := \sqrt{n}(\hat{\Theta}\tilde{X}_1'\tilde{X}_1/n - I)(\hat{\eta} - \eta_0)$$

[89] show that $\|\Delta\|_\infty = o_p(1)$ when λ for the nodewise regression is chosen at rate $\sqrt{\log n/n}$

Notice that from (A)

$$\hat{Q}\lambda\hat{\kappa} = (M_n \circ \hat{D}_n)'(D_n - X_n\hat{\beta} - (M_n \circ \hat{D}_n)\hat{\eta})/n$$

Thus

$$\begin{aligned}
\hat{e} &= \hat{\eta} + \hat{\Theta}(M_n \circ \hat{D}_n)'(D_n - X_n\hat{\beta} - (M_n \circ \hat{D}_n)\hat{\eta})/n \\
&= \eta_0 + \frac{1}{n}\hat{\Theta}\tilde{X}_1'\epsilon_1 - \Delta/\sqrt{n} \\
&\to \eta_0 \quad \text{as } n \to \infty
\end{aligned}$$

Similarly

$$\begin{aligned}
(\hat{\beta} - \beta_0) &= (X_n'X_n)^{-}X_n'\epsilon_1 - (X_n'X_n)^{-}X_n'(M_n \circ \hat{D}_n)(\hat{\eta} - \eta_0) \\
&= (X_n'X_n)^{-}X_n'\left(I - (M_n \circ \hat{D}_n)\hat{\Theta}\tilde{X}_1'/n\right)\epsilon_1 + (X_n'X_n)^{-}X_n'(M_n \circ \hat{D}_n)\Delta/\sqrt{n} \\
&+ (X_n'X_n)^{-}X_n'(M_n \circ \hat{D}_n)\hat{\Theta}\hat{Q}\lambda\hat{\kappa}
\end{aligned}$$

So

$$\begin{aligned}\hat{b} &= \hat{\beta} - (X_n'X_n)^{-}X_n'(M_n \circ \hat{D}_n)\hat{\Theta}(M_n \circ \hat{D}_n)'(D_n - (M_n \circ \hat{D}_n)\hat{\eta} - X_n\hat{\beta})/n \\ &= \beta_0 + (X_n'X_n)^{-}X_n'\left(I - (M_n \circ \hat{D}_n)\hat{\Theta}\tilde{X}_1'/n\right)\epsilon_1 + (X_n'X_n)^{-}X_n'(M_n \circ \hat{D}_n)\Delta/\sqrt{n} \\ &\to \beta_0 \quad \text{as } n \to \infty\end{aligned}$$

Notice that the estimator $\hat{b}$ is a special case in [26]

Now consider the design matrix in the second stage, $M_n \circ \hat{D}_n$. Let $(.)_S$ be the operator that restricts a matrix to its columns indexed in S.

Define $\Sigma^x_{1,1,n} = \frac{1}{n}\left(M_n \circ \hat{D}_n\right)_S'\left(M_n \circ \hat{D}_n\right)_S$.
Define $\Sigma^x_{2,1,n} = \frac{1}{n}\left(M_n \circ \hat{D}_n\right)_{S^c}'\left(M_n \circ \hat{D}_n\right)_S$.

$$\begin{aligned}\Sigma^x_{1,1,n} &= \frac{1}{n} diag\left((\hat{D}_n)_S\right)\left(M_n\right)_S'\left(M_n\right)_S diag\left((\hat{D}_n)_S\right) \\ &= \frac{1}{n} diag\left((\hat{D}_n)_S\right)\Sigma_{1,1,n} diag\left((\hat{D}_n)_S\right)\end{aligned}$$

So

$$\left(\Sigma^x_{1,1,n}\right)^{-1} = n \cdot diag\left((\hat{D}_n)_S\right)^{-1}\Sigma^{-1}_{1,1,n} diag\left((\hat{D}_n)_S\right)^{-1}$$

And,

$$\begin{aligned}\Sigma^x_{2,1,n} &= \frac{1}{n} diag\left((\hat{D}_n)_{S^c}\right)\left(M_n\right)_{S^c}'\left(M_n\right)_S diag\left((\hat{D}_n)_S\right) \\ &= \frac{1}{n} diag\left((\hat{D}_n)_{S^c}\right)\Sigma_{2,1,n} diag\left((\hat{D}_n)_S\right)\end{aligned}$$

Thus

$$\left\|\Sigma^x_{2,1,n}\left(\Sigma^x_{1,1,n}\right)^{-1} sign(\eta_0)\right\|_\infty = \left\|diag\left((\hat{D}_n)_{S^c}\right)\Sigma_{2,1,n}\Sigma^{-1}_{1,1,n} diag\left((\hat{D}_n)_S\right)^{-1} sign(\eta_0)\right\|_\infty$$

Assume $\hat{D}_n \to \Gamma$, $\Sigma_{2,1,n}\Sigma_{1,1,n}^{-1} \to \Sigma$, then I require

$$||diag(\Gamma_S)\Sigma diag(\Gamma_{S^c})sign(\eta_0)||_\infty < 1$$

The consistency of the active set $\lim_{n\to\infty} \mathbb{P}(\hat{S}_n = S) = 1$ follows from [93] under Assumption 5. □

A.1.2 Theorem 2

Proof. In the presence of multiple networks, we use sparse square root lasso:

$$\min_{\eta}\left\{\frac{1}{\sqrt{n}}\left\|D_n-\sum_{j=1}^{q}(M_n^j\circ\hat{D}_n)\eta^j-X_n\beta\right\|_2+\left(\sum_{j=1}^{q}\sqrt{T_j}\Big(\lambda_1\|\eta^j\|_2+\lambda_2\|\eta^j\|_1\Big)\right)\right\}$$

The KKT condition with respect to the jth group can be written as:

$$\frac{-(M_n^j\circ\hat{D}_n)'(D_n-\sum_{j=1}^{q}(M_n^j\circ\hat{D}_n)\hat{\eta}^j-X_n\hat{\beta})}{\sqrt{n}\|D_n-\sum_{j=1}^{q}(M_n^j\circ\hat{D})\hat{\eta}^j-X_n\hat{\beta}\|_2}+\lambda_1\tau^j+\lambda_2\nu^j=0 \tag{A1}$$

For any $\hat{\beta}_i^j\neq 0$ in group j,

$$\tau_i^j=\frac{\sqrt{T_j}\hat{\eta}_i^j}{\|\hat{\eta}^j\|_2},\text{ and }\nu_i^j=\sqrt{T_j}sign(\hat{\eta}_i^j)$$

Let $\tau=(\tau_1,\tau_2,\cdots,\tau_p)'$, $\nu=(\nu_1,\nu_2,\cdots,\nu_p)'$. Let $\hat{Z}_n=\Big[(M_n^1\circ\hat{D}_n),(M_n^2\circ\hat{D}_n),\cdots(M_n^q\circ\hat{D}_n)\Big]$, $Z_n=\Big[(M_n^1\circ D_n),(M_n^2\circ D_n),\cdots(M_n^q\circ D_n)\Big]$. $\hat{Q}:=\|D_n-\sum_{j=1}^{q}(M_n^j\circ\hat{D}_n)\hat{\eta}^j-X_n\hat{\beta}\|_2$. Let $\eta=(\eta^{1'},\eta^{2'},\cdots,\eta^{q'})'$. Plug in $D_n=Z_n\eta_0+X_n\beta_0+\epsilon_n$. (A1) can be transformed as:

$$-\frac{\hat{Z}_n'\epsilon}{n}+\frac{\hat{Z}_n'\hat{Z}_n}{n}(\hat{\eta}-\eta_0)+\frac{\hat{Z}_n'X_n}{n}(\hat{\beta}-\beta_0)+\underbrace{\frac{1}{n}\hat{Z}_n'\Big(\hat{Z}_n-Z_n\Big)\eta_0}_{(C*)}+\sqrt{n}\hat{Q}\lambda_1\tau+\sqrt{n}\hat{Q}\lambda_2\nu=0 \tag{A2}$$

The derivative with respect to β is

$$\begin{aligned}&-X_n'(D_n-X_n\hat{\beta}-\hat{Z}_n\hat{\eta})/n=0\\ \Leftrightarrow&\frac{1}{n}X_n'X_n(\hat{\beta}-\beta_0)+\frac{1}{n}X_n'\hat{Z}(\hat{\eta}-\eta_0)+\frac{1}{n}X_n'\Big(\hat{Z}_n-Z_n\Big)\eta_0=\frac{X_n'\epsilon}{n}\end{aligned} \tag{A3}$$

Notice that $D_n=(I-\sum_{j=1}^{q}M_n^j\circ\eta^j)^{-}X_n\beta+(I-\sum_{j=1}^{q}M_n^j\circ\eta^j)^{-}\epsilon=f(\sum_{j=1}^{q}M_n^j\circ X_n)+\epsilon_1$.

Equation (C*) can be written as:

$$\begin{aligned}\frac{1}{n}\hat{Z}'\Big([M_n^1\circ(\hat{D}_n-D_n),M_n^2\circ(\hat{D}_n-D_n),\cdots,M_n^q\circ(\hat{D}_n-D_n)]\Big)\eta_0\\ =\frac{1}{n}\hat{Z}'\Big([M_n^1\circ(\hat{D}_n-f),M_n^2\circ(\hat{D}_n-f),\cdots,M_n^q\circ(\hat{D}_n-f)]\Big)\eta_0\\ +\frac{1}{n}\hat{Z}'\Big([M_n^1\circ(f-D_n),M_n^2\circ(f-D_n),\cdots,M_n^q\circ(f-D_n)]\Big)\eta_0\end{aligned}$$

By Theorem 3, $\|\hat{D}_n-f\|_2/\sqrt{n}<M\lambda\sqrt{s_n}$. When $\lambda\asymp\sqrt{\frac{\log n}{n}}$, $\|\hat{D}_n-f\|_2=o(n^{1/4})$

By assumption 1* and 5*

$$\begin{aligned}\left\|\frac{1}{n}\hat{Z}'\sum_{j=1}^{q}\Big(M_n^j\circ(\hat{D}_n-f)\Big)\eta_0^j\right\|_\infty &\le\sum_{j=1}^{q}\left\|\frac{1}{n}\hat{Z}'\Big(M_n^j\circ(\hat{D}_n-f)\Big)\eta_0^j\right\|_\infty\\ &\le\sum_{j=1}^{q}\max_{i=1,\cdots,q}\left\|\frac{1}{n}M_n^i(M_n^j\circ\eta_0)(\hat{D}_n-f)\right\|_\infty\|\hat{D}_n\|_\infty\\ &\le\frac{1}{n}\sum_{j=1}^{q}\max_{i=1,\cdots,q}\|M_n^i(M_n^j\circ\eta_0)\|_{op2}\left\|(\hat{D}_n-f)\right\|_2\|\hat{D}_n\|_\infty=o(1/\sqrt{n})\end{aligned}$$

Since

$$\max_{i=1,\cdots,q}\|M_n^i(M_n^j\circ\eta_0)\|_{op2}\le\max_{i=1,\cdots,q}\|M_n^i\|_{op2}\|(M_n^j\circ\eta_0)\|_{op2}=o(n^{1/4})$$

where $\|.\|_{op2}$ is the operator norm from $l_2\to l_\infty$

And

$$\begin{aligned}\frac{1}{n}\hat{Z}'\Big([M_n^1\circ(f-D_n),M_n^2\circ(f-D_n),\cdots,M_n^q\circ(f-D_n)]\Big)\eta_0\\ =-\frac{1}{n}\hat{Z}'\Big([M_n^1\circ\epsilon_1,M_n^2\circ\epsilon_1,\cdots,M_n^q\circ\epsilon_1]\Big)\eta_0\\ =-\frac{1}{n}\hat{Z}'\sum_{j=1}^{q}\Big(M_n^j\circ\eta_0^j\Big)(I-\sum_{j=1}^{q}M_n^j\circ\eta_0^j)^-\epsilon\\ =-\frac{1}{n}\hat{Z}'\left(\Big(I-\sum_{j=1}^{q}M_n^j\circ\eta_0^j\Big)^--I\right)\epsilon\end{aligned}$$

Thus (A2) and (A3) can be written as:

$$\frac{1}{n}\hat{Z}_n'X_n(\hat{\beta}-\beta_0)+\frac{1}{n}\hat{Z}_n'\hat{Z}_n(\hat{\eta}-\eta_0)+\sqrt{n}\hat{Q}\lambda_1\tau+\sqrt{n}\hat{Q}\lambda_2\nu+o(1/\sqrt{n})=\frac{\hat{Z}_n'\epsilon_1}{n}$$
$$\frac{1}{n}X_n'X_n(\hat{\beta}-\beta_0)+\frac{1}{n}X_n'\hat{Z}_n(\hat{\eta}-\eta_0)+o(1/\sqrt{n})=\frac{X_n'\epsilon_1}{n}$$

Define $\tilde{Z}_m = W_n\hat{Z}_n$. Find $\hat{\Theta}_Z$ as an approximation for the inverse of $\tilde{Z}_m'\tilde{Z}_m/n$

$$(\hat{\eta}-\eta_0)+\sqrt{n}\hat{\Theta}_Z Q(\lambda_1\tau+\lambda_2\nu)=\hat{\Theta}_Z\tilde{Z}_m'\epsilon_1/n-\Delta_m/\sqrt{n}$$
$$(\hat{\beta}-\beta_0)-\sqrt{n}(X_n'X_n)^{-1}X_n'\hat{Z}_n\hat{\Theta}_Z Q(\lambda_1\tau+\lambda_2\nu)=(X_n'X_n)^{-1}X_n'\Big(I-\hat{Z}_n\hat{\Theta}_Z\tilde{Z}_m'/n\Big)\epsilon_1$$
$$+(X_n'X_n)^{-1}X_n'\hat{Z}_n\Delta_m/\sqrt{n}$$

where $\Delta_m = \sqrt{n}(\hat{\Theta}_Z\tilde{Z}_m'\tilde{Z}_m/n-I)(\hat{\eta}-\eta_0)$

This suggest the following estimator:

$$\hat{e}_m=\hat{\eta}+\hat{\Theta}_Z\hat{Z}_n'(D_n-\hat{Z}_n\hat{\eta}-X_n\hat{\beta})/n\to\eta_0$$
$$\hat{b}_m=\hat{\beta}-(X_n'X_n)^{-1}X_n'\hat{Z}_n\hat{\Theta}_Z X_n'(D_n-\hat{Z}_n\hat{\eta}-X_n\hat{\beta})/n\to\beta_0$$

Now consider the design matrix in the second stage, $\left[\left(M_n^1\circ\hat{D}_n\right),\cdots,\left(M_n^q\circ\hat{D}_n\right)\right]$. Let $(.)_S$ be the operator that restricts a matrix to its columns indexed in S.

Define $\Sigma^x_{1,1,n}=\frac{1}{n}\left[\left(M_n^1\circ\hat{D}_n\right)_{S_1},\cdots,\left(M_n^q\circ\hat{D}_n\right)_{S_q}\right]'\left[\left(M_n^1\circ\hat{D}_n\right)_{S_1},\cdots,\left(M_n^q\circ\hat{D}_n\right)_{S_q}\right]$.
Define $\tilde{\Sigma}^x_{2,1,n}=\frac{1}{n}\left[\left(\tilde{M}^1_{S_1^c}\circ\hat{D}_n\right),\cdots,\left(\tilde{M}^q_{S_q^c}\circ\hat{D}_n\right)\right]'\left[\left(\tilde{M}^1_{S_1}\hat{D}_n\right),\cdots,\left(\tilde{M}^q_{S_q}\circ\hat{D}_n\right)\right]$.
where $\tilde{M}^j_{S^c}$ is defined as M_n^j with all non-influential individuals columns being replaced with 0s

Notice that

$$\Sigma^x_{1,1,n}=\frac{1}{n}diag\left(\left[(\hat{D}_n)_{S_1},\cdots,(\hat{D}_n)_{S_q}\right]\right)\left[\left(M_n^1\right)_{S_1},\cdots,\left(M_n^q\right)_{S_q}\right]'\left[\left(M_n^1\right)_{S_1},\cdots,\left(M_n^q\right)_{S_q}\right]$$
$$\cdot diag\left(\left[(\hat{D}_n)_{S_1},\cdots,(\hat{D}_n)_{S_q}\right]\right)$$
$$=\frac{1}{n}diag\left(\left[(\hat{D}_n)_{S_1},\cdots,(\hat{D}_n)_{S_q}\right]\right)\Sigma_{1,1,n}diag\left(\left[(\hat{D}_n)_{S_1},\cdots,(\hat{D}_n)_{S_q}\right]\right)$$

So

$$\left(\Sigma^x_{1,1,n}\right)^{-1} = n \cdot diag\left(\left[(\hat{D}_n)_{S_1}, \cdots, (\hat{D}_n)_{S_q}\right]\right)^{-1} \Sigma^{-1}_{1,1,n} diag\left(\left[(\hat{D}_n)_{S_1}, \cdots, (\hat{D}_n)_{S_q}\right]\right)^{-1}$$

And,

$$\begin{aligned}\Sigma^x_{2,1,n} = & \frac{1}{n} diag\left(\left[\hat{D}_n, \cdots, \hat{D}_n\right]\right)\left[\left(\tilde{M}^1_n\right)_{S^c_1}, \cdots, \left(\tilde{M}^q_n\right)_{S^c_q}\right]' \left[\left(M^1_n\right)_{S_1}, \cdots, \left(M^q_n\right)_{S_q}\right] \\ & \cdot diag\left(\left[(\hat{D}_n)_{S_1}, \cdots, (\hat{D}_n)_{S_q}\right]\right) \\ = & \frac{1}{n} diag\left(\left[\hat{D}_n, \cdots, \hat{D}_n\right]\right) \tilde{\Sigma}_{2,1,n} diag\left(\left[(\hat{D}_n)_{S_1}, \cdots, (\hat{D}_n)_{S_q}\right]\right)\end{aligned}$$

Thus

$$\Sigma^x_{2,1,n}\left(\Sigma^x_{1,1,n}\right)^{-1} = diag\left(\left[\hat{D}_n, \cdots, \hat{D}_n\right]\right) \tilde{\Sigma}_{2,1,n} \Sigma^{-1}_{1,1,n} diag\left(\left[(\hat{D}_n)_{S_1}, \cdots, (\hat{D}_n)_{S_q}\right]\right)^{-1}$$

Notice that the jth group in the vector

$$\left(\tilde{\Sigma}^x_{2,1,n}\left(\Sigma^x_{1,1,n}\right)^{-1} u\right)^j = diag(\hat{D}_n)\left(\tilde{\Sigma}_{2,1,n}\Sigma^{-1}_{1,1,n} diag\left(\left[(\hat{D}_n)_{S_1}, \cdots, (\hat{D}_n)_{S_q}\right]\right)^{-1} u\right)^j$$

$$\begin{aligned}& \max_{u:\|u\|_2 \le \sqrt{n}} \max_{1 \le j \le q} \frac{\| \left(\tilde{\Sigma}^x_{2,1,n}\left(\Sigma^x_{1,1,n}\right)^{-1} u\right)^j \|_2}{\sqrt{n}} \\ = & \max_{u:\|u\|_2 \le 1} \max_{1 \le j \le q} \left\| diag(\hat{D}_n)\left(\tilde{\Sigma}_{2,1,n}\Sigma^{-1}_{1,1,n} diag\left(\left[(\hat{D}_n)_{S_1}, \cdots, (\hat{D}_n)_{S_q}\right]\right)^{-1} u\right)^j \right\|_2\end{aligned}$$

The consistency of the active set $\lim_{n\to\infty} \mathbb{P}(\hat{S}_n = S) = 1$ follows Theorem 4. □

A.1.3 Theorem 3

Consider the error term in Theorem 1:

$$\frac{(M_n \circ \hat{D}_n)' W_n (I - M_n \circ \eta_0)^{-1} \epsilon}{n} = diag(\hat{D}_n) \frac{M_n' W_n (I - M_n \circ \eta_0)^{-1} \epsilon}{n}$$

And by assumption,

$$\frac{1}{n} M_n' W_n (I - M_n \circ \eta_0)^{-1} (I - M_n \circ \eta_0)^{-1'} W_n M_n \to \Omega$$

Thus,

$$\frac{M_n' W_n (I - M_n \circ \eta_0)^{-} \epsilon}{\sqrt{n}} \to N(0, \Omega)$$

Notice that the limit exist as

$$\begin{aligned} \left\| \frac{M_n' W_n (I - M_n \circ \eta_0)^{-} \epsilon}{\sqrt{n}} \right\|_\infty &\leq \left\| \frac{M_n' (I - M_n \circ \eta_0)^{-} \epsilon}{\sqrt{n}} \right\|_\infty \|W_n\|_\infty \\ &\leq \frac{1}{\sqrt{n}} \|M_n'\|_{op1} \|(I - M_n \circ \eta_0)^{-}\|_{op1} \|\epsilon\|_1 \|W_n\|_\infty \end{aligned}$$

where $\|.\|_{op1}$ norm is the operation norm from $l_1 \to l_\infty$, which is the maximum entry in the matrix. Notice that

$$\begin{aligned} \|(I - M_n \circ \eta_0)^{-}\|_{op1} &= \| \sum_{k=0}^{\infty} (M_n \circ \eta_0)^k \|_{op1} \\ &\leq \sum_{k=0}^{\infty} \|(M_n \circ \eta_0)\|_{op1}^k \\ &\leq \frac{1}{1 - \eta_{max}} \end{aligned}$$

Also, $\|M_n'\|_{op1} = 1$ and $\|\epsilon\|_1 / \sqrt{n} = O(1)$

As a result

$$\left\| \frac{M_n' W_n (I - M_n \circ \eta_0)^{-} \epsilon}{\sqrt{n}} \right\|_\infty \leq O(1)$$

And the limit exists.

Let $\Gamma = \lim_{n\to\infty} \hat{D}_n$, $\Theta_1 = \lim_{n\to\infty} \hat{\Theta}$, $\hat{Z}_n = (M_n \circ \hat{D}_n)$, $\tilde{Z}_n = X_n(X_n'X_n)^{-1}X_n'\hat{Z}_n$ and $\Theta_2 = \lim_{n\to\infty} \frac{1}{n}\Big(I - \hat{Z}_n\hat{\Theta}\tilde{Z}_n'/n\Big)' X_n(X_n'X_n)^{-1}X_n'\Big(I - \hat{Z}_n\hat{\Theta}\tilde{Z}_n'/n\Big)$

We have:

$$\begin{aligned} \sqrt{n}(\hat{e} - \eta_0) &= E_1 + \Delta_1, \\ \sqrt{n}(\hat{b} - \beta_0) &= E_2 + \Delta_2, \\ E_1 &\sim N(0, \sigma^2\Theta_1 diag(\Gamma)\Omega diag(\Gamma)\Theta_1'), \\ E_2 &\sim N(0, \sigma^2\Theta_2 diag(\Gamma)\Omega diag(\Gamma)\Theta_2'), \end{aligned} \tag{A.1}$$

where $\|\Delta_1\|_\infty = \sqrt{n}(\hat{\Theta}\tilde{X}_n'\tilde{X}_n/n - I)(\hat{\eta} - \eta_0) = o_p(1)$,
and $\|\Delta_2\|_\infty = (X_n'X_n)^{-1}X_n'(M_n \circ \hat{D}_n)\sqrt{n}(\hat{\Theta}\tilde{X}_n'\tilde{X}_n/n - I)(\hat{\eta} - \eta_0) = o_p(1)$

A.1.4 Theorem 4

The error term in Theorem 2:

$$\begin{aligned}\tilde{Z}_n'\epsilon_1/n &= \left[(M_n^1 \circ \hat{D}_n), (M_n^2 \circ \hat{D}_n), \cdots (M_n^q \circ \hat{D}_n)\right]' W_n(I - M_n \circ \eta_0)^{-1}\epsilon/n \\ &= diag(\hat{D}_n)\left[M_n^1, M_n^2, \cdots M_n^q\right]' W_n(I - M_n \circ \eta_0)^{-1}\epsilon/n\end{aligned}$$

and by assumption,

$$\left[M_n^1, M_n^2, \cdots M_n^q\right]' W_n(I - M_n \circ \eta_0)^{-1}\epsilon/\sqrt{n} \to N(0, \Omega_m)$$

Let $\Gamma = \lim_{n\to\infty} \hat{D}_n$, $\Theta_{Z1} = \lim_{n\to\infty} \hat{\Theta}_Z$, $\hat{Z}_n = \left[(M_n^1 \circ \hat{D}_n), (M_n^2 \circ \hat{D}_n), \cdots (M_n^q \circ \hat{D}_n)\right]$, $\tilde{Z}_n = X_n(X_n'X_n)^{-1}X_n'\hat{Z}_n$ and $\Theta_{Z2} = \lim_{n\to\infty} \frac{1}{n}\left(I - \hat{Z}_n\hat{\Theta}_Z\tilde{Z}_n'/n\right)' X_n(X_n'X_n)^{-1}X_n'\left(I - \hat{Z}_n\hat{\Theta}_Z\tilde{Z}_n'/n\right)$

We have:

$$\begin{aligned}\sqrt{n}(\hat{e}_m - \eta_0) &= E_{m1} + \Delta_{m1}, \\ \sqrt{n}(\hat{b}_m - \beta_0) &= E_{m2} + \Delta_{m2}, \\ E_{m1} &\sim N(0, \sigma^2\Theta_{Z1}diag(\Gamma)\Omega_2 diag(\Gamma)\Theta_{Z1}'), \\ E_{m2} &\sim N(0, \sigma^2\Theta_{Z2}diag(\Gamma)\Omega_2 diag(\Gamma)\Theta_{Z2}'),\end{aligned} \tag{A.2}$$

where $\|\Delta_{m1}\|_\infty = \sqrt{n}(\hat{\Theta}_Z\tilde{Z}_n'\tilde{Z}_n/n - I)(\hat{\eta} - \eta_0) = o_p(1)$,
and $\|\Delta_{m2}\|_\infty = (X_n'X_n)^- X_n'\hat{Z}_n \sqrt{n}(\hat{\Theta}_Z\tilde{Z}_n'\tilde{Z}_n/n - I)(\hat{\eta} - \eta_0) = o_p(1)$

A.1.5 Square-root Sparse Group LASSO

To prove Theorem 2 and Theorem 4, we need the following results from square-root sparse group LASSO: 1) Bounds on the prediction, i.e. $\left\|\sum_{j=1}^{q}(M^j \circ X_n)(\hat{\eta}^j - \eta_0^j) + X_n(\hat{\beta} - \beta_0)\right\|_2 \lesssim \lambda$. And 2) Consistency of selection i.e. $\hat{S}_n = S$.

First, define the effective networks as:

$$SG := \{1 \leq j \leq q : |\eta_0^j|_1 \neq 0\}$$

Define the influential individuals in network j as:

$$S^j := \{1 \leq i \leq n : \eta_{0i}^j \neq 0\}$$

Define the number of influential individuals in network j as $|S^j| = s_n^j$. Define the number of the all influential individuals as $|S| = \sum_{j=1}^{q} s_n^j = s_n$. Define the number of non-zero groups as $|SG| = s_g$. Let $\eta_S \in \mathbb{R}^{s_n}$ be the coefficients for the influential individuals and $\eta_{S^c} \in \mathbb{R}^{nq-s_n}$ be the coefficients for the non-influential individuals.

Theorem 12. *Assume $\kappa > 0$, $\gamma > 1$ and $\alpha \in (0,1)$. Assume $\max_j s_n^j \leq \frac{n}{\log n}$ and $s_g \leq \frac{n}{\log q}$. Let $\lambda = \lambda_1 + \lambda_2$. Under assumptions 1*-5*, the following holds with probability greater than $1-\alpha$:*

$$\left\|\sum_{j=1}^{q}(M^j \circ X_n)(\hat{\eta}^j - \eta_0^j) + X_n(\hat{\beta} - \beta_0)\right\|_2 \lesssim \frac{\sigma\lambda\sqrt{ns_n}}{\kappa}, \tag{A.3}$$

and

$$\sum_{j=1}^{q}\sqrt{T_j}\|(\hat{\eta} - \eta_0)^j\|_2 \lesssim \frac{\sigma\lambda s_n}{\kappa} \tag{A.4}$$

Theorem 5 establishes bounds on the LASSO prediction. Notice that Theorem 5 is still valid under a weaker assumption (compatibility assumption) than assumption 4*. The details are shown in the proofs.

The advantage of using sparse group LASSO compare to standard LASSO is that consistent model selection can be achieved under a weaker condition. Standard LASSO requires the l_2 norm of the correlations between all irrelevant regressors and relevant regressors to be small. On the other hand, sparse group LASSO only requires that the l_2 norm of correlations between irrelevant regressors in each group and relevant regressors to be small.

Theorem 13. *Define* $c = \lambda_1/\lambda$. *For constant* $\vartheta < 1$, $\alpha \in (0,1)$ *and* $D > 0$, *if* $c < \frac{1-\vartheta}{2}$ *and under assumptions 1*-5*, with probability greater than* $1-\alpha$:

1. $\hat{\eta}_{S^c} = 0$,
2. *for all* $1 \leq j \leq q$,
$$\|(\hat{\eta} - \eta_0)^j\|_\infty \leq D(c\sqrt{n} + 1 - c)\sigma\lambda,$$
3. *if* $\min\{\eta_0^j\} \geq D\sqrt{T_j}\sigma\lambda$, *then*
$$S = \hat{S}$$

Theorem 6 shows that consistent selection can be achieved if the design matrix satisfies the Irrepresentable condition together with a Beta-min condition. The ratio between λ_2 and λ_1 is $\frac{1}{c} - 1 > \frac{1+\vartheta}{1-\vartheta}$. Thus when the correlation among the irrelevant regressors in each group and relevant regressors is low, we can penalize more on the l_2 norm and *vice versa*.

A.1.6 Theorem 5

In the proof of theorem 3 and 4, I consider the following standard lasso problem:

$$\min_{\beta}\left\{\frac{\|Y-X\beta\|_2}{\sqrt{n}}+\left(\sum_{j=1}^{q}\sqrt{T_j}\Big(\lambda_1\|\eta_j\|_2+\lambda_2\|\eta\|_1\Big)\right)\right\} \tag{A4}$$

where the true data generating process is $Y = X\beta + \sigma\epsilon$, where ϵ s a mean 0 process with variance 1.

I use β^0, σ to represent the true parameter values. Let p be the total number of regressors. Let $\{G_1, \cdots, G_q\}$ be a partition of $\{1, \cdots, p\}$ and $T_i = |G_i|$, $i = 1, \cdots q$ be the number of regressors in each group. Denote $\beta^j \in \mathbb{R}^{T_j}$ as the coefficients for regressors in group j. Both p, q and T_is can go to ∞ as $n \to \infty$. Define the active group as:

$$SG := \{1 \le j \le q : |\beta^{0j}|_1 \neq 0\}$$

Define the active set among all regressors as:

$$S := \{1 \le i \le p : \beta_i^0 \neq 0\}$$

Define the size of true support of β^0 as $|S| = s$; define the number of non-zeros groups as $|SG| = s_g$. Let $\beta_S \in \mathbb{R}^{s_1}$ be the set of coefficients on the true support and $\beta_{S^c} \in \mathbb{R}^{p-s_1}$ be the coefficients for those irrelevant regressors.

Define $\hat{Q}(\beta) := \frac{\|Y-X\beta\|_2^2}{n}$. Define $\hat{\delta} := \hat{\beta} - \beta^0$.

The advantage of using square root type lasso is the tuning parameter λ_1 and λ_2 can be chosen independently from σ. In sparse group lasso, the noise component can be viewed in two different ways:

1. I want λ_1 to be sufficiently large to overrule the noise component in grouped lasso, defined as:

$$V = \max_{1 \le j \le q} \left\{ \frac{\sqrt{n} \|(X'\epsilon)^j\|_2}{\sqrt{T_j} \|\epsilon\|_2} \right\}$$

2. I want λ_2 to be sufficiently large to overrule the noise component in standard lasso within each group, defined as:

$$\Lambda = n \left\| \nabla \sqrt{\hat{Q}(\beta^0)} \right\|_\infty = \max_{1 \le j \le q} \left\{ \frac{\sqrt{n} \|(X'\epsilon)^j\|_\infty}{\|\epsilon\|_2} \right\}$$

Lemma 3. *Assume the noise terms ϵ_i are i.i.d standard normal random variables. Let $\alpha \in (0, 1)$ be given such that $p/\alpha > 8$ and $n > \log(1/\alpha)$. If*

$$\lambda \ge \sqrt{2 \log(4p/\alpha)/n}$$

Then

$$\mathbb{P}(\Lambda \ge n\lambda) \le \alpha/2$$

Lemma 3 is a direct result as case (ii) of Lemma 1 in [13]

Notice that a direct inequality:

$$\mathbb{P}(V \ge n\lambda) \le \mathbb{P}(\Lambda \ge n\lambda)$$

as $\|(X'\epsilon)^j\|_2 / \sqrt{T_j} \le \|(X'\epsilon)^j\|_\infty$.

Define the event $\mathcal{A}_1 := \{V \le n\lambda/\bar{\gamma}\}$, the set $\mathcal{A}_2 := \{\Lambda \le n\lambda/\bar{\gamma}\}$. We can choose

$$\min\{\lambda_1, \lambda_2\} \ge \sqrt{2 \log(4p/\alpha)/n}$$

So that:

$$\mathbb{P}(\mathcal{A} := \mathcal{A}_1 \cap \mathcal{A}_2) \ge \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2) - 1 \ge 1 - \alpha$$

Here, $\bar{\gamma} = \frac{\gamma+1}{\gamma-1}$, where γ is defined as below.

Define

$$\Delta_\gamma^1 := \{\delta \in \mathbb{R}^p : \sum_{j \in SG^c} \sqrt{T_j}\|\delta^j\|_2 \leq \gamma \sum_{j \in SG} \sqrt{T_j}\|\delta^j\|_2\}$$

$$\Delta_\gamma^2 := \{\delta \in \mathbb{R}^p : \sum_{j \in SG^c} \sqrt{T_j}\|\delta_{S^c}^j\|_1 \leq \gamma \sum_{j \in SG} \sqrt{T_j}\|\delta_S^j\|_1\}$$

Compatibility Condition (CC). *We say that the Compatibility Condition is met for* $\kappa > 0$ *and* $\gamma > 1$ *if :*

$$\sum_{j \in SG} \sqrt{T_j}\|\hat{\delta}_S^j\|_1 \leq \frac{\sqrt{s_n}\|X\hat{\delta}\|_2}{\sqrt{n}\kappa}$$

for all $\delta \in \Delta_\gamma^1 \cap \Delta_\gamma^2$

Proof. • First, by definition of (A4)

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta^0)} \leq \underbrace{\lambda_1 \sum_{j=1}^q \sqrt{T_j}(\|\beta^{0j}\|_2 - \|\hat{\beta}^j\|_2)}_{(1)} + \overbrace{\lambda_2 \sum_{j=1}^q \sqrt{T_j}(\|\beta^{0j}\|_1 - \|\hat{\beta}^j\|_1)}^{(2)} \tag{A5}$$

$$\begin{aligned}
(1) &= \lambda_1 \sum_{j \in SG} \sqrt{T_j}(\|\beta^{0j}\|_2 - \|\hat{\beta}^j\|_2) - \lambda_1 \sum_{j \in SG^c} \sqrt{T_j}(\|\hat{\beta}^j\|_2) \\
&\leq \lambda_1 \sum_{j \in SG} \sqrt{T_j}(|\|\beta^{0j}\|_2 - \|\hat{\beta}^j\|_2|) - \lambda_1 \sum_{j \in SG^c} \sqrt{T_j}(\|\hat{\beta}^j\|_2) \\
&\leq \lambda_1 \sum_{j \in SG} \sqrt{T_j}(\|\hat{\delta}^j\|_2) - \lambda_1 \sum_{j \in SG^c} \sqrt{T_j}(\|\hat{\delta}^j\|_2)
\end{aligned}$$

$$\begin{aligned}
(2) &= \lambda_2 \sum_{j=1}^q \sqrt{T_j}(\|\beta_S^{0j}\|_1 - \|\hat{\beta}_S^j\|_1 - \|\hat{\beta}_{S^c}^j\|_1) \\
&\leq \lambda_2 \sum_{j=1}^q \sqrt{T_j}(\|\hat{\delta}_S^j\|_1 - \|\hat{\delta}_{S^c}^j\|_1)
\end{aligned}$$

- Second, by convexity,

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta^0)} \geq \nabla \sqrt{\hat{Q}(\beta^0)}(\hat{\delta}) \geq -\frac{|\epsilon' X\hat{\delta}|}{\sqrt{n}\|\epsilon\|_2}$$

$$\begin{aligned}
|\epsilon' X\hat{\delta}| &= \Big| \sum_{j=1}^{q} \epsilon' X^j \hat{\delta}^j \Big| \leq \sum_{j=1}^{q} \|(\epsilon' X^j)'\|_2 \|\hat{\delta}^j\|_2 \\
&\leq \max_{1\leq j\leq q} \left\{ \frac{\sqrt{n}\|(\epsilon' X^j)'\|_2}{\sqrt{T_j}\|\epsilon\|_2} \right\} \frac{\|\epsilon\|_2}{\sqrt{n}} \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j\|_2 \\
&= V \frac{\|\epsilon\|_2}{\sqrt{n}} \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j\|_2
\end{aligned}$$

Also

$$\begin{aligned}
|\epsilon' X\hat{\delta}| &= \Big| \sum_{j=1}^{q} \epsilon' X^j \hat{\delta}^j \Big| \leq \sum_{j=1}^{q} \|(\epsilon' X^j)'\|_\infty \|\hat{\delta}^j\|_1 \\
&\leq \left\| \frac{\sqrt{n}\|(\epsilon' X^j)'\|_\infty}{\sqrt{T_j}\|\epsilon\|_2} \right\|_\infty \frac{\|\epsilon\|_2}{\sqrt{n}} \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j\|_1 \\
&= \frac{\Lambda}{\sqrt{T_j}} \frac{\|\epsilon\|_2}{\sqrt{n}} \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j\|_1
\end{aligned}$$

On set $\mathcal{A}$, we have $\lambda/\bar{\gamma} \geq V$, Thus

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta^0)} \geq -\frac{\lambda}{\bar{\gamma}} \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j\|_2 \tag{A6}$$

Again, on set $\mathcal{A}$, we also have $\lambda/\bar{\gamma} \geq \Lambda/\sqrt{T_{min}}$, Thus

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta^0)} \geq -\frac{\lambda}{\bar{\gamma}} \sum_{j=1}^{q} \sqrt{T_j}(\|\hat{\delta}^j_S\|_1 + \|\hat{\delta}^j_{S^c}\|_1) \tag{A7}$$

- Third, Combine (A6) and (A7), for any $c \in [0, 1]$

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta^0)} \geq -\frac{c\lambda}{\bar{\gamma}} \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j\|_2 - \frac{(1-c)\lambda}{\bar{\gamma}} \sum_{j=1}^{q} \sqrt{T_j}(\|\hat{\delta}^j_S\|_1 + \|\hat{\delta}^j_{S^c}\|_1) \tag{A8}$$

Set $\lambda_1 = c\lambda$ and $\lambda_2 = (1-c)\lambda$, we can combine (A8) and (A5) to get

$$
\begin{aligned}
&c\lambda \sum_{j \in SG} \sqrt{T_j}(\|\hat{\delta}^j\|_2) - c\lambda \sum_{j \in SG^c} \sqrt{T_j}(\|\hat{\delta}^j\|_2) + (1-c)\lambda \sum_{j=1}^{q} \sqrt{T_j}(\|\hat{\delta}^j_S\|_1 - \|\hat{\delta}^j_{S^c}\|_1) \\
&\geq -\frac{c\lambda}{\bar{\gamma}} \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j\|_2 - \frac{(1-c)\lambda}{\bar{\gamma}} \sum_{j=1}^{q} \sqrt{T_j}(\|\hat{\delta}^j_S\|_1 + \|\hat{\delta}^j_{S^c}\|_1)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\Big(1 + \frac{1}{\bar{\gamma}}\Big)c\lambda \sum_{j \in SG} \sqrt{T_j}\|\hat{\delta}^j\|_2 + \Big(1 + \frac{1}{\bar{\gamma}}\Big)(1-c)\lambda \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j_S\|_1 \\
&\geq \Big(1 - \frac{1}{\bar{\gamma}}\Big)c\lambda \sum_{j \in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_2 + \Big(1 - \frac{1}{\bar{\gamma}}\Big)(1-c)\lambda \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j_{S^c}\|_1
\end{aligned}
$$

which implies:

$$
\begin{aligned}
&c\gamma \sum_{j \in SG} \sqrt{T_j}\|\hat{\delta}^j\|_2 + (1-c)\gamma \sum_{j \in SG^c} \sqrt{T_j}\|\hat{\delta}^j_S\|_1 \\
&\geq c \sum_{j \in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_2 + (1-c) \sum_{j \in SG^c} \sqrt{T_j}\|\hat{\delta}^j_{S^c}\|_1
\end{aligned}
\tag{A9}
$$

(A9)$\Rightarrow \hat{\delta} \in \Delta^1_\gamma \cap \Delta^2_\gamma$. Thus,

$$
\sum_{j \in SG} \sqrt{T_j}\|\hat{\delta}^j\|_2 \leq \sum_{j \in SG} \sqrt{T_j}\|\hat{\delta}^j_S\|_1 \leq \frac{\sqrt{s_n}\|X\hat{\delta}\|_2}{\sqrt{n}\kappa} \tag{A10}
$$

- Forth, from (A10),

$$
\begin{aligned}
\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta^0)} &\leq \lambda_1 \sum_{j \in SG} \sqrt{T_j}(\|\hat{\delta}^j\|_2) - \lambda_1 \sum_{j \in SG^c} \sqrt{T_j}(\|\hat{\delta}^j\|_2) \\
&\quad + \lambda_2 \sum_{j=1}^{q} \sqrt{T_j}(\|\hat{\delta}^j_S\|_1 - \|\hat{\delta}^j_{S^c}\|_1) \\
&\leq c\lambda \sum_{j \in SG} \sqrt{T_j}(\|\hat{\delta}^j\|_2) + (1-c)\lambda \sum_{j=1}^{q} \sqrt{T_j}\|\hat{\delta}^j_S\|_1 \\
&\leq \lambda \frac{\sqrt{s_n}\|X\delta\|_2}{\sqrt{n}\kappa}
\end{aligned}
\tag{A11}
$$

- Fifth,

$$\begin{aligned}
\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) &= \frac{1}{n}\Big((Y - X\hat{\beta})'(Y - X\hat{\beta}) - (Y - X\beta_0)'(Y - X\beta_0)\Big) \\
&= \frac{1}{n}\{(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
&\quad - (Y - X\hat{\beta} + X\hat{\beta} - X\beta_0)'(Y - X\beta_0)\} \\
&= \frac{1}{n}\{(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
&\quad - (Y - X\hat{\beta})'(Y - X\beta_0) - X(\hat{\beta} - \beta_0)'(Y - X\beta_0)\} \\
&= \frac{1}{n}\{-(Y - X\hat{\beta})'X(\hat{\beta} - \beta_0) - X(\hat{\beta} - \beta_0)'(Y - X\beta_0)\} \\
&= \frac{1}{n}\{(\hat{\beta} - \beta_0)'X'X(\hat{\beta} - \beta_0) - 2X(\hat{\beta} - \beta_0)'(Y - X\beta_0)\} \\
&= \frac{||X\hat{\delta}||_2^2}{n} - \frac{2\sigma\epsilon' X\hat{\delta}}{n}
\end{aligned}$$

- Sixth, from (A11),

$$\begin{aligned}
\frac{||X\hat{\delta}||_2^2}{n} &= \hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) + \frac{2\sigma\epsilon' X\hat{\delta}}{n} \\
&= \Big(\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta^0)}\Big)\Big(\sqrt{\hat{Q}(\hat{\beta})} + \sqrt{\hat{Q}(\beta^0)}\Big) + \frac{2\sigma\epsilon' X\hat{\delta}}{n} \\
&\leq \lambda\frac{\sqrt{s_n}||X\delta||_2}{\sqrt{n}\kappa}\Big(2\sqrt{\hat{Q}(\beta^0)} + \lambda\frac{\sqrt{s_n}||X\delta||_2}{\sqrt{n}\kappa}\Big) + 2V\frac{||\sigma\epsilon||_2}{n^{3/2}}\sum_{j=1}^{q}\sqrt{T_j}||\hat{\delta}^j||_2
\end{aligned}$$

From (A10):

$$\leq \frac{s_n\lambda}{\kappa^2 n}\frac{||X\delta||_2^2}{n} + 2\lambda\frac{||\sigma\epsilon||_2}{\sqrt{n}}\frac{\sqrt{s_n}||X\delta||_2}{\sqrt{n}\kappa} + 2V\frac{||\sigma\epsilon||_2}{n^{3/2}}\frac{\sqrt{s_n}||X\hat{\delta}||_2}{\sqrt{n}\kappa}$$

$n\lambda/\bar{\gamma} \geq V$:

$$\leq \frac{s_n\lambda}{\kappa^2 n}\frac{||X\delta||_2^2}{n} + 2\left(1 + \frac{1}{\bar{\gamma}}\right)\frac{||\sigma\epsilon||_2}{\sqrt{n}}\lambda\frac{\sqrt{s_n}||X\hat{\delta}||_2}{\sqrt{n}\kappa}$$

As a result:

$$\left(1 - \left(\frac{\lambda\sqrt{s_1}}{\kappa}\right)^2\right)\frac{||X\hat{\delta}||_2^2}{n} \leq 2\left(1 + \frac{1}{\bar{\gamma}}\right)\frac{||\sigma\epsilon||_2}{\sqrt{n}}\lambda\frac{\sqrt{s_1}||X\hat{\delta}||_2}{\sqrt{n}\kappa} \tag{A12}$$

(A12) concludes the first statement in Theorem 3.

For the second claim, use the fact that $\delta \in \Delta_\gamma^1$ and the Compatibility Condition:

$$\begin{aligned}\sum_{j=1}^{q} \sqrt{T_j}\|\delta^j\|_2 \leq (\gamma+1) \sum_{j \in SG^c} \sqrt{T_j}\|\delta^j\|_2 &\leq \frac{(\gamma+1)\sqrt{s_n}\|X\delta\|_2}{\sqrt{n}\kappa} \\ &\lesssim \frac{(\gamma+1)\sqrt{s_n}}{\sqrt{n}\kappa} \frac{\sqrt{n}\sigma\lambda\sqrt{s_n}}{\kappa} \lesssim \frac{\sigma\lambda s_n}{\kappa}\end{aligned}$$

□

Lemma 4. *Assumption* (4) implies Irrepresentable Condition: for* $0 < \vartheta < 1$ *if* $\Sigma_{1,1}$ *is invertible and*

$$\max_{u:\|u\|_\infty \le \sqrt{T_k}} \max_{1\le j\le q} \frac{\| \left(\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right)^j \|_\infty}{\sqrt{T_j}} \le \vartheta$$

for all k.

Proof.

$$\|v\|_2 \le \sqrt{T_k} \Rightarrow \|v\|_\infty \le \sqrt{T_k}$$

Thus, Assumption* (4) $\Rightarrow$

$$\max_{u:\|u\|_\infty \le \sqrt{T_k}} \max_{1\le j\le q} \frac{\| \left(\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right)^j \|_2}{\sqrt{T_j}} \le \vartheta$$

Since

$$\| \left(\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right)^j \|_2 \ge \| \left(\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right)^j \|_\infty$$

Thus

$$\max_{u:\|u\|_\infty \le \sqrt{T_k}} \max_{1\le j\le q} \frac{\| \left(\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right)^j \|_\infty}{\sqrt{T_j}} \le \vartheta$$

A.1.7 Theorem 6

From Lemma 2 and Lemma 3, Define $\mathcal{B}_1 = \{V \le n\lambda/(\bar{\gamma} \vee 2\bar{\vartheta})\}$ and $\mathcal{B}_2 = \{\Lambda \le \sqrt{T_{min}}\lambda/(\bar{\gamma} \vee 2\bar{\vartheta})\}$. I can choose

$$\min\{\lambda_1, \lambda_2\} \ge \max\left\{(\bar{\gamma} \vee 2\vartheta)\sqrt{2\log(4p/\alpha)/n}, \frac{\sqrt{n}(\bar{\gamma} \vee 2\vartheta)}{\sqrt{T_{min}}}\sqrt{2\log(4p/\alpha)/n}\right\}$$

So that:

$$\mathbb{P}(\mathcal{B} := \mathcal{B}_1 \cap \mathcal{B}_2) \ge \mathbb{P}(\mathcal{B}_1) + \mathbb{P}(\mathcal{B}_2) - 1 \ge 1 - \alpha \tag{A13}$$

Here, $\bar{\vartheta} = \frac{1+\vartheta}{1-\vartheta}$, where ϑ is defined assumption 4*.

Proof. Choose λ big enough so that (A13) holds.

- First take the derivative of (A4) with respect to each column i:

$$\frac{-(X_i^j)'(Y - X\hat{\beta})}{\|Y - X\hat{\beta}\|_2} = \sqrt{n}\lambda_1\tau_i^j + \sqrt{n}\lambda_2\nu_i^j \tag{A14}$$

Let $\tau = (\tau_1, \tau_2, \cdots, \tau_p)'$, $\nu = (\nu_1, \nu_2, \cdots, \nu_p)'$ $\hat{Q} := \|Y - X\hat{\beta}\|_2$

For any $\hat{\beta}_i^j \neq 0$ in group j,

$$\tau_i^j = \frac{\sqrt{T_j}\hat{\beta}_i^j}{\|\hat{\beta}^j\|_2}, \text{ and } \nu_i^j = \sqrt{T_j}sign(\hat{\beta}_i^j)$$

By KKT condition, $\|\tau^j\|_2 \leq \sqrt{T_j}$ and $|\nu_i^j| \leq \sqrt{T_j}$.

$Y = X\beta + \epsilon$. Let $\hat{delta} = \hat{\beta} - \beta^0$. We can rewrite (A14) in matrix form:

$$\sigma X'\epsilon - X'X\hat{\delta} = \sqrt{n}\hat{Q}\lambda_1\tau + \sqrt{n}\hat{Q}\lambda_2\nu \tag{A15}$$

or

$$\begin{pmatrix} \sigma(X'\epsilon)_S \\ \sigma(X'\epsilon)_{S^c} \end{pmatrix} - n\begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}\begin{pmatrix} \hat{\delta}_S \\ \hat{\delta}_{S^c} \end{pmatrix} = \begin{pmatrix} (\sqrt{n}\hat{Q}\lambda_1\tau + \sqrt{n}\hat{Q}\lambda_2\nu)_S \\ (\sqrt{n}\hat{Q}\lambda_1\tau + \sqrt{n}\hat{Q}\lambda_2\nu)_{S^c} \end{pmatrix}$$

- Second, the upper part of (A15) can be transform to

$$-n\Sigma_{1,1}\hat{\delta}_S - n\Sigma_{1,2}\hat{\delta}_{S^c} = \sqrt{n}\hat{Q}(\lambda_1\tau_S + \lambda_2\nu_S) - \sigma(X'\epsilon)_S$$

or, equivalently,

$$\begin{aligned} &-n\hat{\delta}'_{S^c}\Sigma_{2,1}\hat{\delta}_S - n\hat{\delta}'_{S^c}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}\hat{\delta}_{S^c} \\ &\quad = \sqrt{n}\hat{\hat{Q}}\hat{\delta}'_{S^c}\Sigma_{2,1}\Sigma_{1,1}^{-1}(\lambda_1\tau_S + \lambda_2\nu_S) - \sigma\hat{\delta}'_{S^c}\Sigma_{2,1}\Sigma_{1,1}^{-1}(X'\epsilon)_S \end{aligned} \tag{A16}$$

Notice that for all $\hat{\delta}_i^j \neq 0$ but $i \in S^c$, either $j \in SG^c$ or $j \in SG$.

Define

$$SG_1 \subset SG := \{1 \leq j \leq q : \exists \beta_i^{0j} = 0\}$$

Define

$$S^{jc} := \{1 \leq i \leq T_j : \beta_i^{0j} = 0\}$$

Let $l_j = |S^{jc}|$ denotes the size of the sparsity in group j.

The right hand side of (A16) can be broken into two parts. The first part consider all sparse term in nonzero groups while the second term consider all zero groups:

$$(A16) = \underbrace{\sqrt{n}\hat{Q}\lambda \sum_{j \in SG_1} \sum_{i \in S^{jc}} \hat{\delta}_i^j \left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\Big(c\tau_S + (1-c)\nu_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\Big) \right]_i^j}_{(1)}$$
$$+ \underbrace{\sqrt{n}\hat{Q}\lambda \sum_{j \in SG^c} \sum_{i \in S^{jc}} \hat{\delta}_i^j \left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\Big(c\tau_S + (1-c)\nu_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\Big) \right]_i^j}_{(2)}$$

$$(1) = \underbrace{\sqrt{n}\hat{Q}\lambda c \sum_{j \in SG_1} \sum_{i \in S^{jc}} \hat{\delta}_i^j \left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\Big(\tau_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\Big) \right]_i^j}_{(3)}$$
$$+ \underbrace{\sqrt{n}\hat{Q}\lambda(1-c) \sum_{j \in SG_1} \sum_{i \in S^{jc}} \hat{\delta}_i^j \left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\Big(\nu_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\Big) \right]_i^j}_{(4)}$$

$$(2) = \underbrace{\sqrt{n}\hat{Q}\lambda c \sum_{j \in SG^c} \sum_{i \in S^{jc}} \hat{\delta}_i^j \left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\Big(\tau_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\Big) \right]_i^j}_{(5)}$$
$$+ \underbrace{\sqrt{n}\hat{Q}\lambda(1-c) \sum_{j \in SG^c} \sum_{i \in S^{jc}} \hat{\delta}_i^j \left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\Big(\nu_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\Big) \right]_i^j}_{(6)}$$

By Holder:

$$(3) \leq \sqrt{n}\hat{Q}\lambda c \sum_{j\in SG_1}\left\{\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\left(\tau_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\right)\right]_i^j\right\|_\infty\right\}$$

Observe again that if $n\lambda/2\bar{\vartheta} \geq \hat{\Lambda}/\sqrt{T_{min}} \Rightarrow \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}((X_i^j)'\epsilon) \leq \frac{\sqrt{T_j}}{2\bar{\vartheta}}$ for any i and

$$\|\tau^j\|_\infty = \left\|\frac{\sqrt{T_j}\hat{\beta}_i^j}{\|\hat{\beta}^j\|_2}\right\|_\infty \leq \sqrt{T_j}$$

By Lemma 4

$$\begin{aligned}(3) &\leq \sqrt{n}\hat{Q}\lambda c \max_{u:\|u\|_\infty\leq\left(\sqrt{T_j}+\frac{\sqrt{T_j}}{2\bar{\vartheta}}\right)}\sum_{j\in SG_1}\left\{\left(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\right)\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right]^j\right\|_\infty\right\}\\ &\leq \left(1+\frac{1}{2\bar{\vartheta}}\right)\sqrt{n}\hat{Q}\lambda c \max_{u:\|u\|_\infty\leq\sqrt{T_j}}\sum_{j\in SG_1}\sqrt{T_j}\left\{\left(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\right)\frac{\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right]^j\right\|_\infty}{\sqrt{T_j}}\right\}\\ &\leq \vartheta\left(1+\frac{1}{2\bar{\vartheta}}\right)\sqrt{n}\hat{Q}\lambda c\sum_{j\in SG_1}\sqrt{T_j}\left(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\right)\end{aligned}$$

By Holder:

$$(4) \leq \sqrt{n}\hat{Q}\lambda(1-c)\sum_{j\in SG_1}\left\{\left(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\right)\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\left(\nu_S - \frac{\sqrt{n}\sigma}{\lambda\hat{Q}}(X'\epsilon)_S\right)\right]^j\right\|_\infty\right\}$$

Observe that if $\lambda/2\bar{\vartheta} \geq \hat{\Lambda}/\sqrt{T_{min}} \Rightarrow \frac{\sqrt{n}\sigma}{\lambda\hat{Q}}((X^i)'\epsilon) \leq \frac{\sqrt{T_j}}{2\bar{\vartheta}}$ for any i and $\|\nu_i^j\|_\infty \leq \sqrt{T_j}$.

$$\begin{aligned}(4) &\leq \sqrt{n}\hat{Q}\lambda(1-c)\max_{u:\|u\|_\infty\leq\left(1+\frac{1}{2\bar{\vartheta}}\right)\sqrt{T_k}}\sum_{j\in SG_1}\left\{\left(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\right)\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right]^j\right\|_\infty\right\}\\ &\leq \left(1+\frac{1}{2\bar{\vartheta}}\right)\sqrt{n}\hat{Q}\lambda(1-c)\max_{u:\|u\|_\infty\leq\sqrt{T_k}}\sum_{j\in SG_1}\sqrt{T_j}\left\{\left(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\right)\frac{\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}u\right]^j\right\|_\infty}{\sqrt{T_j}}\right\}\\ &\leq \vartheta\left(1+\frac{1}{2\bar{\vartheta}}\right)\sqrt{n}\hat{Q}\lambda(1-c)\sum_{j\in SG_1}\sqrt{T_j}\left\{\left(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\right)\right\}\end{aligned}$$

Since $\|\tau^j\|_2 \leq \sqrt{T_j}$ and $n\lambda/\bar{\vartheta} \geq \lambda/2\bar{\vartheta} \geq \hat{V} \Rightarrow \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\|(X'\epsilon)^j\|_2 \leq \frac{\sqrt{T_j}}{\bar{\vartheta}}$

$$
\begin{aligned}
(5) &\leq \sqrt{n}\hat{Q}\lambda c \sum_{j\in SG^c} \sqrt{T_j}\left(\sum_{i\in S^{jc}}(\hat{\delta}_i^j)^2\right)^{1/2} \frac{\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\left(\tau_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\right)\right]^j\right\|_2}{\sqrt{T_j}} \\
&\leq \sqrt{n}\hat{Q}\lambda c \max_{v:\|v^k\|_2\leq(1+\frac{1}{\vartheta})\sqrt{T_k}} \sum_{j\in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_2 \frac{\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}v\right]^j\right\|_2}{\sqrt{T_j}} \\
&\leq \vartheta\left(1+\frac{1}{\bar{\vartheta}}\right) c\sqrt{n}\hat{Q}\lambda \sum_{j\in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_2 \\
&= \left(1-\frac{1}{\bar{\vartheta}}\right) c\sqrt{n}\hat{Q}\lambda \sum_{j\in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_2
\end{aligned}
$$

For any $i \in S$, $v_i^j = \sqrt{T_j}sign(\beta_i^j)$. Thus $|v_i^j| \leq \sqrt{T_j}$. $n\lambda/\bar{\vartheta} \geq n\lambda/2\bar{\vartheta} \geq \hat{\Lambda}/\sqrt{T_{min}} \Rightarrow \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\|(X'\epsilon)^j\|_\infty \leq \frac{\sqrt{T_j}}{\bar{\vartheta}}$

$$
\begin{aligned}
(6) &\leq \sqrt{n}\hat{Q}\lambda(1-c) \sum_{j\in SG^c} \|\hat{\delta}^j\|_1 \left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}\left(v_S - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_S\right)\right]^j\right\|_\infty \\
&\leq \sqrt{n}\hat{Q}\lambda c \max_{v:\|v^k\|_2\leq(1+\frac{1}{\vartheta})\sqrt{T_k}} \sum_{j\in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_1 \frac{\left\|\left[\tilde{\Sigma}_{2,1}\Sigma_{1,1}^{-1}v\right]^j\right\|_\infty}{\sqrt{T_j}} \\
&\leq \vartheta\left(1+\frac{1}{\bar{\vartheta}}\right)(1-c)\sqrt{n}\hat{Q}\lambda \sum_{j\in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_1 \\
&= \left(1-\frac{1}{\bar{\vartheta}}\right)(1-c)\sqrt{n}\hat{Q}\lambda \sum_{j\in SG^c} \sqrt{T_j}\|\hat{\delta}^j\|_1
\end{aligned}
$$

- Third, the bottom part of (A15) can be transform to:

$$
-n\Sigma_{2,1}\hat{\delta}_S - n\Sigma_{2,2}\hat{\delta}_{S^c} = \sqrt{n}\hat{Q}(\lambda_1\tau_{S^c} + \lambda_2 v_{S^c}) - \sigma(X'\epsilon)_{S^c}
$$

or, equivalently,

$$
\begin{aligned}
&-n\hat{\delta}'_{S^c}\Sigma_{2,1}\hat{\delta}_S - n\hat{\delta}'_{S^c}\Sigma_{2,2}\hat{\delta}_{S^c} \\
&\qquad = \sqrt{n}\hat{Q}\hat{\delta}'_{S^c}(\lambda_1\tau_{S^c} + \lambda_2 v_{S^c}) - \sigma\hat{\delta}'_{S^c}(X'\epsilon)_{S^c}
\end{aligned} \tag{A17}
$$

For $j \in SG_1$ and $i \in S^{jc}$,

$$\hat{\beta}_i^j \neq 0 \Rightarrow \hat{\delta}_i^j(c\tau_i^j + (1-c)\nu_i^j) = c\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\beta}^j||_2} + (1-c)\sqrt{T_j}|\delta_i^j|$$

$$\hat{\beta}_i^j = 0 \Rightarrow \hat{\delta}_i^j(c\tau_i^j + (1-c)\nu_i^j) = 0 = c\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\beta}^j||_2} + (1-c)\sqrt{T_j}|\delta_i^j|$$

For $j \in SG^c$ and all i,

$$\hat{\beta}_i^j \neq 0 \Rightarrow \hat{\delta}_i^j(c\tau_i^j + (1-c)\nu_i^j) = c\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\delta}^j||_2} + (1-c)\sqrt{T_j}|\delta_i^j|$$

$$\hat{\beta}_i^j = 0 \Rightarrow \hat{\delta}_i^j(c\tau_i^j + (1-c)\nu_i^j) = 0 = c\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\delta}^j||_2} + (1-c)\sqrt{T_j}|\delta_i^j|$$

As in the previous section, the right hand side of (A17) can be broken into two parts:

$$(A18) = \underbrace{\sqrt{n}\hat{Q}\lambda \sum_{j\in SG_1}\sum_{i\in S^{jc}}\Big(c\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\beta}^j||_2} + (1-c)\sqrt{T_j}|\delta_i^j| - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big)}_{(7)}$$
$$+ \underbrace{\sqrt{n}\hat{Q}\lambda \sum_{j\in SG^c}\sum_{i\in S^{jc}}\Big(c\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\delta}^j||_2} + (1-c)\sqrt{T_j}|\delta_i^j| - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big)}_{(8)}$$

$$(7) = \underbrace{\sqrt{n}\hat{Q}\lambda c \sum_{j\in SG_1}\sum_{i\in S^{jc}}\Big(\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\beta}^j||_2} - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big)}_{(9)}$$
$$+ \underbrace{\sqrt{n}\hat{Q}\lambda(1-c) \sum_{j\in SG_1}\sum_{i\in S^{jc}}\Big(\sqrt{T_j}|\delta_i^j| - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big)}_{(10)}$$

$$(8) = \underbrace{\sqrt{n}\hat{Q}\lambda c \sum_{j\in SG^c}\sum_{i\in S^{jc}}\Big(\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{||\hat{\delta}^j||_2} - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big)}_{(11)}$$
$$+ \underbrace{\sqrt{n}\hat{Q}\lambda(1-c) \sum_{j\in SG^c}\sum_{i\in S^{jc}}\Big(\sqrt{T_j}|\delta_i^j| - \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big)}_{(12)}$$

By Holder, we have

$$
\begin{aligned}
(9) = \ & \sqrt{n}\hat{Q}\lambda c \sum_{j\in SG_1}\Big(\sum_{i\in S^{jc}}\frac{\sqrt{T_j}(\hat{\delta}_i^j)^2}{\|\hat{\beta}^j\|_2} - \sum_{i\in S^{jc}}\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big) \\
\geq \ & \sqrt{n}\hat{Q}\lambda c \sum_{j\in SG_1}\left\{\frac{\sum_{i\in S^{jc}}\sqrt{T_j}(\hat{\delta}_i^j)^2}{\|\hat{\beta}^j\|_2} - \Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big)\left\|\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_i^j\right\|_\infty\right\}
\end{aligned}
$$

Observe again that if $n\lambda/2\bar{\vartheta} \geq \hat{\Lambda}/\sqrt{T_{min}} \Rightarrow \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}((X^i)'\epsilon) \leq \frac{\sqrt{T_j}}{2\bar{\vartheta}}$ for any i

$$
\begin{aligned}
(9) \geq \ & \sqrt{n}\hat{Q}\lambda c \sum_{j\in SG_1}\left\{\frac{\sqrt{T_j}\sum_{i\in S^{jc}}(\hat{\delta}_i^j)^2}{\|\hat{\beta}^j\|_2} - \frac{\sqrt{T_j}}{2\bar{\vartheta}}\Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big)\right\} \\
\geq \ & -\frac{1}{2\bar{\vartheta}}\sqrt{n}\hat{Q}\lambda c \sum_{j\in SG_1}\sqrt{T_j}\Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big)
\end{aligned}
$$

$$
\begin{aligned}
(10) = \ & \sqrt{n}\hat{Q}\lambda(1-c)\left\{\sqrt{T_j}\sum_{j\in SG_1}\Big(\sum_{i\in S^{jc}}|\delta_i^j|\Big) - \Big(\sum_{i\in S^{jc}}\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big)\right\} \\
\geq \ & \sqrt{n}\hat{Q}\lambda(1-c)\sum_{j\in SG_1}\left\{\sqrt{T_j}\Big(\sum_{i\in S^{jc}}|\delta_i^j|\Big) - \Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big)\left\|\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}(X'\epsilon)_i^j\right\|_\infty\right\} \\
\geq \ & \left(1-\frac{1}{2\bar{\vartheta}}\right)\sqrt{n}\hat{Q}\lambda(1-c)\sum_{j\in SG_1}\sqrt{T_j}\Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big)
\end{aligned}
$$

Since $n\lambda/\bar{\vartheta} \geq \lambda/2\bar{\vartheta} \geq \hat{V} \Rightarrow \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\|(X'\epsilon)^j\|_2 \leq \frac{\sqrt{T_j}}{\bar{\vartheta}}$ for any j:

$$
\begin{aligned}
(11) = \ & \sqrt{n}\hat{Q}\lambda c \sum_{j\in SG^c}\Big(\sqrt{T_j}\|\hat{\delta}^j\|_2 - \sum_{i\in S^{jc}}\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big) \\
\geq \ & \sqrt{n}\hat{Q}\lambda c \sum_{j\in SG^c}\Big(\sqrt{T_j}\|\hat{\delta}^j\|_2 - \sqrt{T_j}\|\hat{\delta}^j\|_2\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\frac{\|(X'\epsilon)^j\|_2}{\sqrt{T_j}}\Big) \\
\geq \ & \left(1-\frac{1}{\bar{\vartheta}}\right)c\sqrt{n}\hat{Q}\lambda\sum_{j\in SG^c}\sqrt{T_j}\|\hat{\delta}^j\|_2
\end{aligned}
$$

Since $n\lambda/\bar{\vartheta} \geq \hat{\Lambda}/\sqrt{T_{min}} \Rightarrow \frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\|(X'\epsilon)^j\|_\infty \leq \frac{\sqrt{T_j}}{\bar{\vartheta}}$ for any j:

$$
\begin{aligned}
(12) = \ & \sqrt{n}\hat{Q}\lambda(1-c)\sum_{j\in SG^c}\Big(\sqrt{T_j}\|\hat{\delta}^j\|_1 - \sum_{i\in S^{jc}}\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\hat{\delta}_i^j(X'\epsilon)_i^j\Big) \\
\geq \ & \sqrt{n}\hat{Q}\lambda(1-c)\sum_{j\in SG^c}\Big(\sqrt{T_j}\|\hat{\delta}^j\|_1 - \|\hat{\delta}^j\|_1\frac{\sigma}{\sqrt{n}\lambda\hat{Q}}\|(X'\epsilon)^j\|_\infty\Big) \\
\geq \ & \left(1-\frac{1}{\bar{\vartheta}}\right)(1-c)\sqrt{n}\hat{Q}\lambda\sum_{j\in SG^c}\sqrt{T_j}\|\hat{\delta}^j\|_1
\end{aligned}
$$

Subtract (A17) from (A16) and notice (11) and (12) canceled with (5) and (6), we have:

$$
\begin{aligned}
& n^2\hat{\delta}'_{S^c}(\Sigma_{2,2}-\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})\hat{\delta}_{S^c} \\
& \leq \left\{-\left(1-\frac{1}{2\bar{\vartheta}}\right)(1-c)+\frac{1}{2\bar{\vartheta}}c+\vartheta\Big(1+\frac{1}{2\bar{\vartheta}}\Big)c+\vartheta\Big(1+\frac{1}{2\bar{\vartheta}}\Big)(1-c)\right\} \\
& \cdot\ \sqrt{n}\hat{Q}\lambda\sum_{j\in SG_1}\Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big) \\
& = \left\{-1+\frac{1}{2\bar{\vartheta}}+\vartheta\Big(1+\frac{1}{2\bar{\vartheta}}\Big)+c+\vartheta\Big(1+\frac{1}{2\bar{\vartheta}}\Big)c-\vartheta\Big(1+\frac{1}{2\bar{\vartheta}}\Big)c\right\} \\
& \cdot\ \sqrt{n}\hat{Q}\lambda\sum_{j\in SG_1}\Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big) \\
& = \left\{-\frac{1}{2}(1-\vartheta)+c\right\}\sqrt{n}\hat{Q}\lambda\sum_{j\in SG_1}\Big(\sum_{i\in S^{jc}}|\hat{\delta}_i^j|\Big) \\
& \leq 0
\end{aligned}
$$

The last inequality is due to Substitution Condition.

However, since $\Sigma_{2,2}-\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2} \geq 0$, this implies $\hat{\delta}_{S^c} = 0$, which establish the first claim.

For the second claim, substitute $\hat{\delta}_{S^c} = 0$ into (A16) we have:

$$
-n\hat{\delta}_S = \sqrt{n}\hat{Q}\lambda\Sigma_{1,1}^{-1}\left(c\tau_S + (1-c)\nu_S - \frac{\sigma(X'\epsilon)_S}{\sqrt{n}\hat{Q}\lambda}\right)
$$

Recall again $n\lambda/\bar{\vartheta} \geq \hat{\Lambda}/\sqrt{T_{min}}$

$$
\begin{aligned}
\|\hat{\delta}^j\|_\infty &\leq \frac{\hat{Q}\lambda c}{n^{1/2}} \left\| \left[\tilde{\Sigma}_{1,1}^{-1} \left(\tau_S - \frac{\sigma(X'\epsilon)_S}{\sqrt{n}\hat{Q}\lambda} \right) \right]^j \right\|_\infty \\
&+ \frac{\hat{Q}\lambda(1-c)}{n^{1/2}} \left\| \left[\tilde{\Sigma}_{1,1}^{-1} \left(\nu_S - \frac{\sigma(X'\epsilon)_S}{\sqrt{n}\hat{Q}\lambda} \right) \right]^j \right\|_\infty \\
&\leq \frac{\hat{Q}\lambda c}{n^{1/2}} \max_{v:\|v^k\|_2 \leq (1+\frac{1}{\vartheta})\sqrt{T_j}} \left\| \left[\tilde{\Sigma}_{1,1}^{-1} v \right]^j \right\|_\infty \\
&+ \frac{\hat{Q}\lambda(1-c)}{n^{1/2}} \max_{u:\|u\|_\infty \leq (1+\frac{1}{\vartheta})\sqrt{T_j}} \left\| \left[\tilde{\Sigma}_{1,1}^{-1} u \right]^j \right\|_\infty \\
&\leq (1 + \frac{1}{\bar{\vartheta}}) \frac{\hat{Q}\lambda}{n^{1/2}} \sqrt{T_j} \max_{u:\|u\|_\infty \leq \sqrt{T_j}} \frac{\left\| \left[\tilde{\Sigma}_{1,1}^{-1} u \right]^j \right\|_\infty}{\sqrt{T_j}} \\
&\leq D\sqrt{T_j}\sigma\lambda
\end{aligned}
$$

The third claim follows with Beta Min Condition.

□

A.2 Useful Algebra Transformation

A.2.1 (4)

Since $\Big(M \circ D\Big)\eta = \Big(M \circ \eta\Big)D$,

$$
\begin{aligned}
D_n &= \Big(M_n \circ D_n\Big)\eta_0 + X_n\beta_0 + \epsilon_n \\
\Leftrightarrow D_n &= \Big(M_n \circ \eta_0\Big)D_n + X_n\beta_0 + \epsilon_n \\
\Leftrightarrow &\Big(I_n - \Big(M_n \circ \eta_0\Big)\Big) D_n = X_n\beta_0 + \epsilon_n \\
\Leftrightarrow D_n &= \Big(I_n - \Big(M_n \circ \eta_0\Big)\Big)^{-} (X_n\beta_0 + \epsilon_n) \\
\Leftrightarrow D_n &= \sum_{i=0}^{\infty} \Big(M_n \circ \eta_0\Big)^i (X_n\beta_0 + \epsilon_n)
\end{aligned}
$$

A.2.2 (5)

$$
\begin{aligned}
E(D_n) &= \sum_{i=0}^{\infty} \Big(M_n \circ \eta_0\Big)^i \beta_0 X_n \\
&= \beta_0 X_n + \beta_0\Big(M_n \circ \eta_0\Big)X_n + \sum_{i=2}^{\infty} \Big(M_n \circ \eta_0\Big)^i \beta_0 X_n \\
&= X_n\beta_0 + \Big(M_n \circ X_n\Big)(\beta_0\eta_0) + \sum_{i=2}^{\infty} \Big(M_n \circ \eta_0\Big)^i \beta_0 X_n
\end{aligned}
$$

A.2.3 (6)

Let $M_n = (m_1, m_2, \cdots, m_n)$, where m^j is the jth column of M_n. $\eta_0 = (\eta_1, \eta_2, \cdots, \eta_n)'$

Then

$$\begin{aligned}\Big(M_n \circ \eta_0\Big)^2 &= \Big(M_n \circ \eta_0\Big)\Big(m_1\eta_1, m_2\eta_2, \cdots, m_n\eta_n\Big) \\ &= \Big[\Big(M_n \circ \eta_0\Big)m_1\eta_1, \Big(M_n \circ \eta_0\Big)m_2\eta_2, \cdots, \Big(M_n \circ \eta_0\Big)m_n\eta_n\Big] \\ &= \Big[\Big(M_n \circ m_1\Big)\eta_0\eta_1, \Big(M_n \circ m_2\Big)\eta_0\eta_2, \cdots, \Big(M_n \circ m_n\Big)\eta_0\eta_n\Big]\end{aligned}$$

Thus

$$\begin{aligned}\Big(M_n \circ \eta_0\Big)^2\beta_0 X_n &= \Big(M_n \circ m_1\Big)\eta_0\eta_1 x_{n1}\beta_0 + \Big(M_n \circ m_2\Big)\eta_0\eta_2 x_{n2}\beta_0 + \cdots + \Big(M_n \circ m_n\Big)\eta_0\eta_n x_{nn}\beta_0 \\ &= \Big(M_n \circ m_1\Big)\eta_0\delta_1^1 + \Big(M_n \circ m_2\Big)\eta_0\delta_2^1 + \cdots + \Big(M_n \circ m_n\Big)\eta_0\delta_n^1\end{aligned}$$

$$\begin{aligned}\Big(M_n \circ \eta_0\Big)^3\beta_0 X_n &= \sum_{i=1}^{n}\Big(M_n \circ \eta_0\Big)\Big(M_n \circ m_i\Big)\eta_0\delta_i^1 \\ &= \sum_{i=1}^{n}\Big(M_n \circ \eta_0\Big)\Big(m_1 m_{i1}\eta_1\delta_i^1 + m_2 m_{i2}\eta_2\delta_i^1 + \cdots + m_n m_{in}\eta_n\delta_i^1\Big) \\ &= \sum_{i=1}^{n} m_{i1}\Big(M_n \circ m_1\Big)\eta_0\delta_i^1 + m_{i2}\Big(M_n \circ m_2\Big)\eta_0\delta_i^1 + \cdots + m_{in}\Big(M_n \circ m_n\Big)\eta_0\delta_i^1 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(M_n \circ m_j\Big)\eta_0 m_{ij}\delta_i^1 \\ &= \sum_{i=1}^{n}\Big(M_n \circ m_i\Big)\eta_0\delta_i^2\end{aligned}$$

With induction, one can show that

$$\Big(M_n \circ \eta_0\Big)^k\beta_0 X_n = \sum_{i=1}^{n}\Big(M_n \circ m_i\Big)\eta_0\delta_i^{k-1}$$

Thus,

$$E(D_n|X) = X_n\beta_0 + \Big(M_n \circ X_n\Big)(\beta_0\eta_0) + \sum_{i=1}^{n}\Big(M_n \circ m_i\Big)\eta_0\delta_i^{\infty}$$

where $\delta_i^{\infty} = \sum_{j=1}^{\infty} \delta_i^j$.

When M_n is the adjacency matrix, the jth column of $\left(M_n \circ m_i\right)$ is 0 if $m_{ij} = 0$ or i is not connect with j; and is equal to m_j if $m_{ij} = 1$

Thus

$$\begin{aligned}
E(D_n) &= X_n\beta_0 + \left(m_1 x_1 \eta_1 \beta_0 + m_2 x_2 \eta_2 \beta_0 + \cdots + m_n x_n \eta_n \beta_0\right) \\
&+ \sum_{i=1}^{n} \left(m_1 m_{i1} \eta_1 \delta_i^{\infty} + m_2 m_{i2} \eta_2 \delta_i^{\infty} + \cdots + m_n m_{in} \eta_n \delta_i^{\infty}\right) \\
&= X_n\beta_0 + \left(m_1 x_1 \eta_1 \Big(\beta_0 + \sum_{i=1}^{n} \frac{m_{i1}\delta_i^{\infty}}{x_1}\Big) + m_2 x_2 \eta_2 \Big(\beta_0 + \sum_{i=1}^{n} \frac{m_{i2}\delta_i^{\infty}}{x_2}\Big) + \cdots \right. \\
&\left. + m_n x_n \eta_n \Big(\beta_0 + \sum_{i=1}^{n} \frac{m_{in}\delta_i^{\infty}}{x_n}\Big)\right) \\
&= X_n\beta_0 + \left(M_n \circ X_n\right)\tilde{\eta}
\end{aligned}$$

where $\tilde{\eta}_j = \eta_j\Big(\beta_0 + \sum_{i=1}^{n} \frac{m_{ij}\delta_i^{\infty}}{x_j}\Big)$.

As a result, using $\left(M_n \circ X_n\right)$ and X_n are sufficient to determine the influential individuals.

A.2.4 (8)

$$\begin{aligned}
D_n &= \left(M_n \circ D_n\right)\eta_0 + \gamma M_n D_n + X_n\beta_0 + \epsilon_n \\
\Leftrightarrow D_n &= \left(M_n \circ \eta_0\right)D_n + \gamma M_n D_n + X_n\beta_0 + \epsilon_n \\
\Leftrightarrow &\left(I_n - \left(M_n \circ \eta_0\right) - \gamma M_n\right) D_n = X_n\beta_0 + \epsilon_n \\
\Leftrightarrow D_n &= \left(I_n - \left(M_n \circ \eta_0\right) - \gamma M_n\right)^{-} (X_n\beta_0 + \epsilon_n) \\
\Leftrightarrow D_n &= \sum_{i=0}^{\infty} \left(M_n \circ \eta_0 + \gamma M_n\right)^i (X_n\beta_0 + \epsilon_n)
\end{aligned}$$

A.3 Multiple Networks Assumptions

Assumption* 1. *Among n individuals in q_n networks, let S_n^j be the set of influential individuals in network j. Let $s_n^j = |S_n^j|$ be the number of elements in S_n^j.*

$$s_n^j = o\left(\frac{\sqrt{n}}{\log n}\right), \quad as\ n \to \infty$$
$$s_g = \sum_{j=1}^{q_n} \mathbf{1}(s_n^j \neq 0) = o\left(\frac{n}{\log q_n}\right), \quad as\ n \to \infty$$

Notice same individual from different networks are counted as different elements in S_n

Assumption* 2.

- *There exists an $\eta_{\max} < 1$ such that $\sum_{j=1}^{q} \|\eta_0^j\|_\infty \leq \eta_{max}$*
- *The ϵ_j are i.i.d with 0 mean and variance σ^2*
- *The regressors x_i in X_n are uniformly bounded constants for all n. $\lim_{n\to\infty} X_n'X_n/n$ exists and is nonsingular*

Apply the same algebra:

$$
\begin{aligned}
D_n &= \sum_{j=1}^{q} \Big(M_n^j \circ D_n\Big)\eta_0^j + X_n\beta_0 + \epsilon_n \\
\Leftrightarrow D_n &= \sum_{j=1}^{q} \Big(M_n^j \circ \eta_0^j\Big)D_n + X_n\beta_0 + \epsilon_n \\
\Leftrightarrow &\left(I - \sum_{j=1}^{q} \Big(M_n^j \circ \eta_0^j\Big)\right) D_n = X_n\beta_0 + \epsilon_n \\
\Leftrightarrow D_n &= \left(I - \sum_{j=1}^{q} \Big(M_n^j \circ \eta_0^j\Big)\right)^{-} (X_n\beta_0 + \epsilon_n) \\
\Leftrightarrow D_n &= \sum_{i=0}^{\infty} \Big(\sum_{j=1}^{q} M_n^j \circ \eta_0^j\Big)^i (X_n\beta_0 + \epsilon_n)
\end{aligned}
$$

Consider X_n as a one dimensional vector:

$$
\begin{aligned}
E(D_n) &= \sum_{i=0}^{\infty} \Big(\sum_{j=1}^{q} M_n^j \circ \eta_0^j\Big)^i \beta_0 X_n \\
&= \beta_0 X_n + \beta_0\Big(\sum_{j=1}^{q} M_n^j \circ \eta_0^j\Big) X_n + \sum_{i=2}^{\infty} \Big(\sum_{j=1}^{q} M_n^j \circ \eta_0^j\Big)^i \beta_0 X_n \\
&= X_n\beta_0 + \sum_{j=1}^{q} \Big(M_n^j \circ X_n\Big)(\beta_0\eta_0^j) + \sum_{i=2}^{\infty} \Big(\sum_{j=1}^{q} M_n^j \circ \eta_0^j\Big)^i \beta_0 X_n
\end{aligned}
$$

$X_n, \Big(M_n^1 \circ X_n\Big), \Big(M_n^2 \circ X_n\Big), \cdots, \Big(M_n^q \circ X_n\Big)$ are valid instruments.

Apply the same algebra

$$
E(D_n) = X_n\beta_0 + \sum_{j=1}^{q} \Big(M_n^j \circ X_n\Big)\tilde{\eta}^j
$$

Assumption* 3. $\Big[X_n, \Big(M_n^1 \circ X_n\Big)_S, \Big(M_n^2 \circ X_n\Big)_S, \cdots, \Big(M_n^q \circ X_n\Big)_S\Big]$ *is full rank with probability equals to 1.*

We can use Group Lasso to identify those influential individuals and the networks that deliver the influence. For Group Lasso to achieve consistent selection, we need the following assumption:

Assumption* 4.

(Group Irrepresentable Condition) There exists $N \in \mathbb{N}$*:* $\forall n \geq N$*, there is a* $\vartheta \in (0,1)$ *such that*

$$P\left(\max_{u:\|u\|_2\leq 1}\max_{1\leq j\leq q}\left\|diag(\hat{D}_n)\left(\tilde{\Sigma}_{2,1,n}\Sigma_{1,1,n}^{-1}diag\left(\left[(\hat{D}_n)_{S_1},\cdots,(\hat{D}_n)_{S_q}\right]\right)^{-1}u\right)^j\right\|_2 \leq \vartheta\right) = 1$$

(Beta Min Condition) There exists $N \in \mathbb{N}$*:* $\forall n \geq N$*, there is a* $m > 0$ *such that*

$$\min|(\eta_0)_S| \geq m/\sqrt{n}$$

Assumption* 5.

(Maximum Neighbors Condition)

$$\|M_n^{j'}\mathbf{1}_n\|_\infty \leq O(\log n) \quad \textit{for all } j$$

(Variance Condition)

$$\frac{1}{n}M_n^{0\prime}W_n(I-\sum_{j=1}^{q}M_n^j\circ\eta_0^j)^{-1}(I-\sum_{j=1}^{q}M_n^j\circ\eta_0^j)^{-1\prime}W_nM_n^0 \to \Omega_2$$

where $M_n^0 = \left[M_n^1, M_n^2, \cdots M_n^q\right]$*, and* $W_n = \left(I - X_n(X_n'X_n)^{-1}X_n'\right)$

Define $\Sigma_{1,1,n} = \frac{1}{n}\left[\left(M_n^1\right)_S,\cdots,\left(M_n^q\right)_S\right]'\left[\left(M_n^1\right)_S,\cdots,\left(M_n^q\right)_S\right]$

Define $\Sigma_{2,1,n} = \frac{1}{n}\left[\left(M_n^1\right)_{S^c},\cdots,\left(M_n^q\right)_{S^c}\right]'\left[\left(M_n^1\right)_S,\cdots,\left(M_n^q\right)_S\right]$.

Define $\tilde{\Sigma}_{2,1,n} = \frac{1}{n}\left[\left(\tilde{M}_{S^c}^1\right),\cdots,\left(\tilde{M}_{S^c}^q\right)\right]'\left[\left(\tilde{M}_S^1\right),\cdots,\left(\tilde{M}_S^q\right)\right]$.

where $\tilde{M}_{S^c}^j$ is defined as M_n^j with all non-influential individuals columns being replaced with 0s

A.4 Adjacency Matrix for Influential Individuals

We use the following adjacency matrix for influential individuals when there are five of them:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

We use the following adjacency matrix for influential individuals when there are ten of them:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A.5 Centraility

Denote a graph as $G = (V, E)$, where V represents the set for vertex and E represents the set for edges. Define a measure $d : (x, y) \to \mathbb{R}$ as the length of the shortest path between the node x and y. And define M as the adjacency matrix for graph G. I consider the following centrality measures:

- Degree centrality

  The degree centrality $C_D(x)$ of a vertex V is defined as the number of edges connected to node v.

- Closeness centrality

  The Closeness centrality is the average length of the shortest path between the node and all other nodes in the graph.

  $$C_C(v) = \frac{1}{\sum_{y \in V} d(v, y)}$$

- Betweenness centrality

  Betweenness centrality measures the number of times a node acts as a bridge along the shortest path between two other nodes.

  $$C_B(v) = \sum_{x \neq v \neq y \in V} \frac{\sigma_{x,y}(v)}{\sigma_{x,y}}$$

  where $\sigma_{x,y}$ is total number of shortest paths from node x to node y. $\sigma_{x,y}(v)$ is the number of those paths that pass through v.

- Eigenvector centrality

  Eigenvector centrality is defined as the left-hand eigenvector of the adjacency matrix M associated with the largest eigenvalue λ:

  $$\lambda x = xM$$

  And the vth entry of x is the eigenvector centrality of v.

A.6 Tables

Table A.1: Simulation

	0.1			0.2		
Network Size	50	200	500	50	200	500
Avgcov S_0	0.9780	0.9560	0.9380	0.9770	0.9480	0.9580
Avglength S_0	2.9420	3.6734	2.6136	1.0179	3.3098	2.0386
Avgcov S_0^c	0.9222	0.9861	0.9846	0.9920	0.9861	0.9884
Avglength S_0^c	18.9664	8.1006	2.5444	21.4923	3.1052	1.9782
Avgcov β	0.8700	0.9700	0.9650	0.9500	0.9650	0.9800
Avglength β	4.0056	0.4890	0.2959	0.9773	0.7905	0.5209
Power [1]	0.2140	0.1800	0.4650	0.5870	0.2520	0.1770
FDR [2]	0.0147	0.0001	0.0000	0.0017	0.0000	0.0030
Avgcov 1	0.9900	0.9000	0.9900	0.9500	0.9000	0.9850
Avglength 1	3.0138	5.6446	0.8600	0.7202	4.8020	1.9351
Avgcov 2	0.9700	0.9700	0.8600	0.9850	0.9650	0.9550
Avglength 2	5.0537	2.6632	1.5617	1.5222	2.7718	2.5142
Avgcov 3	0.9800	0.9400	0.9150	0.9900	0.9250	0.9900
Avglength 3	2.4503	4.8645	4.3329	0.7604	4.2660	1.5686
Avgcov 4	0.9600	0.9800	0.9650	0.9950	0.9500	0.9950
Avglength 4	1.8805	3.6298	4.1772	0.8212	3.6354	1.6983
Avgcov 5	0.9900	0.9900	0.9600	0.9650	1.0000	0.8650
Avglength 5	2.3115	1.5647	2.0417	1.2652	1.0741	2.4768

This table summarizes the results simulated on Erdos-Renyi type random graphs. When a node is added into the graph, it has probability $p = 0.1$ or $p = 0.2$ to form a link with all existing nodes.

The reported coverage is from 200 simulations from 50, 200 and 500 nodes graphs. The active set S_0 contains 5 nodes and coverage for each is reported as Avgcov 1-5. Nonactive set S_0^c contains all remaining nodes.

1. Power represents the averaged percentage in the active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

2. FDR reports the averaged percentage in the non-active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

Table A.2: Simulation

	0.1			0.2		
Network Size	50	200	500 [3]	50	200	500 [3]
Avgcov S_0	0.9730	0.9870	0.8805	0.6905	0.9870	0.8330
Avglength S_0	11.8263	1.5870	4.5802	0.8400	3.6207	2.1104
Avgcov S_0^c	0.9942	0.9905	0.9638	0.9827	0.9972	0.9733
Avglength S_0^c	23.2425	2.5128	4.0562	9.3871	2.9423	5.5000
Avgcov β	0.9800	0.9700	0.9300	0.9500	0.9950	0.9950
Avglength β	2.6520	0.5203	0.9008	1.2524	0.5261	0.7915
Power [1]	0.0475	0.1680	0.5725	0.8175	0.1620	0.4710
FDR [2]	0.0000	0.0001	0.0000	0.0102	0.0000	0.0004
Avgcov 1	0.9250	1.0000	0.9200	0.8150	0.9950	0.8800
Avglength 1	6.8760	0.9064	1.6038	0.5357	0.8415	2.0727
Avgcov 2	0.9600	1.0000	0.9350	0.8350	1.0000	0.8550
Avglength 2	5.4525	1.0753	17.1486	0.3364	0.8152	1.2254
Avgcov 3	0.9950	0.9900	0.9300	0.9750	0.9650	0.6600
Avglength 3	12.2042	2.8843	2.5877	2.0093	1.8027	1.4287
Avgcov 4	0.9600	0.9600	0.9450	0.4150	0.9750	0.8300
Avglength 4	17.9822	1.8353	1.5234	0.3987	2.1970	1.7792
Avgcov 5	0.9900	1.0000	0.8600	1.0000	1.0000	0.8150
Avglength 5	3.0831	1.0112	1.1228	0.4478	1.0124	0.7712
Avgcov 6	0.9750	0.9900	0.8850	0.5250	1.0000	0.9100
Avglength 6	39.1903	2.1300	14.2821	0.3061	20.5070	3.9918
Avgcov 7	0.9750	0.9600	0.8500	0.2800	0.9600	0.8250
Avglength 7	19.0146	2.5613	3.2858	0.8466	5.3259	5.0870
Avgcov 8	0.9650	1.0000	0.9100	0.7800	0.9850	0.7950
Avglength 8	3.2928	1.2513	1.2913	0.3709	1.4046	1.3949
Avgcov 9	0.9900	0.9700	0.9250	0.9250	0.9950	0.9500
Avglength 9	3.8329	1.1019	1.4152	2.9014	1.4500	1.2169
Avgcov 10	0.9950	1.0000	0.6450	0.3550	0.9950	0.8100
Avglength 10	7.3340	1.1136	1.5418	0.2474	0.8831	2.1366

This table summarizes the results simulated on Erdos-Renyi type random graphs. When a node is added into the graph, it has probability $p = 0.1$ or $p = 0.2$ to form a link with all existing nodes.

The reported coverage is from 200 simulations from 50, 200 and 500 nodes graphs. The active set S_0 contains 10 nodes and coverage for each is reported as Avgcov 1-10. Nonactive set S_0^c contains all remaining nodes.

1. Power represents the averaged percentage in the active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

2. FDR reports the averaged percentage in the non-active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

3. For 500 cases, lasso tuning parameter is chosen using rule of thumb instead of cross-validation

Table A.3: Simulation: small world

	0.04			0.08		
Network Size	50	200	500	50	200	500
Avgcov S_0	0.9180	0.8490	0.9920	0.9410	0.8310	0.9860
Avglength S_0	5.7298	1.6333	1.6646	5.2069	3.8309	0.9132
Avgcov S_0^c	0.9543	0.9646	0.9809	0.9577	0.9581	0.9949
Avglength S_0^c	7.8860	5.2748	4.3686	3.4985	2.9435	3.4044
Avgcov β	0.9900	0.9350	0.9933	0.9350	0.9650	0.9950
Avglength β	0.8524	0.4044	0.9067	0.7532	0.5382	1.4130
Power [1]	0.0340	0.4350	0.1013	0.1640	0.1020	0.5470
FDR [2]	0.0026	0.0000	0.0056	0.0039	0.0000	0.0000
Avgcov 1	0.8450	0.7650	1.0000	0.9700	0.7950	1.0000
Avglength 1	22.9085	1.1822	2.7360	12.8441	1.1567	0.5102
Avgcov 2	0.9550	0.7200	0.9933	0.9400	0.8400	1.0000
Avglength 2	1.6373	5.2736	2.2481	10.5694	8.0125	0.5399
Avgcov 3	0.9150	0.8900	0.9933	0.8850	0.8100	1.0000
Avglength 3	1.3111	0.6948	2.0804	0.7724	0.8774	0.5530
Avgcov 4	0.9500	0.8900	0.9933	0.9550	0.8600	1.0000
Avglength 4	1.1413	0.3973	0.6932	1.0133	1.9001	0.5196
Avgcov 5	0.9250	0.9800	0.9800	0.9550	0.8500	0.9300
Avglength 5	1.6509	0.6189	0.5653	0.8354	7.2078	2.4434

This table summarizes the results simulated on small-world type random graphs. Given the number of node $N = 50, 200, 500$, the mean degree for each node is $0.04N$ and $0.08N$. The rewriting probability is fixed at 0.4.

The reported coverage is for 200 simulations from 50, 200 and 500 nodes graphs. The active set S_0 contains 5 nodes and coverage for each is reported as Avgcov 1-5. Nonactive set S_0^c contains all remaining nodes.

1. Power represents the averaged percentage in the active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

2. FDR reports the averaged percentage in the non-active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

Table A.4: Simulation

	0.1			0.2		
Network Size	50	200	500[5]	50	200	500[5]
Avgcov S_0	0.9860	0.9940	0.9990	0.8950	0.9910	1.0000
Avglength S_0	10.2325	0.6168	1.4046	6.1020	0.7531	1.1056
Avgcov S_0^c	0.9923	0.9884	0.9868	0.9893	0.9909	0.9945
Avglength S_0^c	9.2108	2.2820	4.0344	5.9378	1.7050	1.9574
Avgcov β	0.9900	0.9650	1.0000	0.9750	0.9650	0.9900
Avglength β	8.2338	0.5556	1.1923	6.4205	0.6026	1.0265
Power [1]	0.3480	0.6810	0.2300	0.5480	0.5620	0.1340
FDR [2]	0.0037	0.0003	0.0001	0.0076	0.0026	0.0019
Network 1:						
probability[3]	0.8050	0.8950	0.3700	0.7400	0.8500	0.2300
# identifed [4]	2.3540	3.8547	3.2973	4.0743	3.3314	2.9778
Network 2						
probability [3]	0.0450	0.0550	0.0300	0.1300	0.0350	0.0100
# identifed [4]	4.3333	1.0000	2.1667	3.4615	1.0000	1.0000
Avgcov 1	0.9750	0.9950	1.0000	0.8850	0.9950	1.0000
Avglength 1	17.4022	0.7528	0.9309	17.1760	0.9289	1.1231
Avgcov 2	0.9950	1.0000	1.0000	0.9450	0.9950	1.0000
Avglength 2	6.7053	0.4484	1.6877	1.7011	0.5777	1.3113
Avgcov 3	0.9900	0.9800	1.0000	0.9750	0.9850	1.0000
Avglength 3	15.5009	0.7919	1.3359	7.9879	1.0755	0.9475
Avgcov 4	0.9800	0.9950	1.0000	0.8250	0.9850	1.0000
Avglength 4	9.6077	0.5766	1.3279	2.7826	0.7049	0.9092
Avgcov 5	0.9900	1.0000	0.9950	0.8450	0.9950	1.0000
Avglength 5	1.9465	0.4742	1.7408	0.8622	0.4783	1.2387

This table summarizes the results simulated on two Erdos-Renyi type random graphs. One of the network (Network 1) passes the endogenous effects while the other one (Network 2) is irrelevant to the decision.

The reported coverage is for 200 simulations from 50, 200 and 500 nodes graphs. The active set S_0 contains 5 nodes and coverage for each is reported as Avgcov 1-5. Nonactive set S_0^c contains all remaining nodes.

1. Power represents the averaged percentage in the active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

2. FDR reports the averaged percentage in the non-active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

3. Probability reports the empirical probability that at least one regressor in the group is significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

4. # identified reports the averaged number of significant regressors in the group conditioning on at least one regressor in the group is significant. False discover rate is controlled at 5% using Benjamini-Hochberg method.

5. For 500 cases, lasso tuning parameter is chosen using rule of thumb instead of cross-validation

Table A.5: Simulation

	0.1			0.2		
Network Size	50	200	500[5]	50	200	500[5]
Avgcov S_0	0.9670	0.9580	0.9850	0.9610	0.9954	0.9980
Avglength S_0	20.3014	1.3383	1.9988	8.6764	2.0044	4.5572
Avgcov S_0^c	0.9665	0.9883	0.9975	0.9680	0.9926	0.9980
Avglength S_0^c	14.0695	3.4002	4.7511	40.5927	1.6113	4.7505
Avgcov β	0.9800	0.9950	0.9900	0.9750	0.9943	0.9950
Avglength β	2.9138	0.8404	0.5866	1.5054	0.6253	0.6881
Avgcov γ	0.9600	0.9950	0.9950	0.9950	1.0000	1.0000
Avglength γ	0.5683	0.1568	0.0257	0.4235	0.0544	0.0294
test-$\gamma \neq 0$	0.4300	0.3750	1.0000	0.4950	1.0000	1.0000
Power [1]	0.0110	0.1890	0.4680	0.0140	0.7726	0.2550
FDR [2]	0.0000	0.0035	0.0000	0.0000	0.0009	0.0000
Avgcov 1	0.9650	0.8100	0.9950	0.9700	1.0000	1.0000
Avglength 1	53.2561	3.1993	1.5312	9.2207	0.2910	5.4525
Avgcov 2	0.9350	0.9950	0.9900	0.9150	0.9886	0.9950
Avglength 2	34.8486	0.6602	0.8785	19.9796	8.4948	0.4050
Avgcov 3	1.0000	0.9950	1.0000	0.9500	1.0000	0.9950
Avglength 3	5.2305	0.9718	4.2919	3.9235	0.3581	3.8538
Avgcov 4	0.9800	0.9950	0.9950	0.9800	0.9943	1.0000
Avglength 4	4.0808	1.0082	2.8069	2.9452	0.5204	2.2894
Avgcov 5	0.9550	0.9950	0.9450	0.9900	0.9943	1.0000
Avglength 5	4.0909	0.8519	0.4855	7.3132	0.3577	10.7855

This table summarizes the results for Heterogeneous Endogenous Effects Model with Cliques.

The reported coverage is for 200 simulations from 50, 200 and 500 nodes graphs. The active set S_0 contains 5 nodes and coverage for each is reported as Avgcov 1-5. Nonactive set S_0^c contains all remaining nodes.

1. Power represents the averaged percentage in the active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

2. FDR reports the averaged percentage in the non-active set being significant after controlling False discover rate at 5% using Benjamini-Hochberg method.

5. For 500 cases, lasso tuning parameter is chosen using rule of thumb instead of cross-validation

Table A.6: Descriptive Statistics

village	number of households	number of villagers	average age	average family size	household having electric	household having latrine	average rooms per person
Village1	182	843	32.7	4.6	90.7%	21.4%	0.6
Village2	195	877	31.4	4.5	94.4%	41.5%	0.5
Village3	294	1384	30.8	4.7	96.9%	47.3%	0.6
Village4	239	1026	31.3	4.3	98.3%	39.7%	0.5
Village12	175	802	30.7	4.6	90.9%	37.7%	0.6
Village19	204	1134	30.9	5.6	87.3%	14.7%	0.3
Village20	156	716	32.8	4.6	80.8%	25.0%	0.4
Village21	202	1046	28.6	5.2	83.7%	16.8%	0.4
Village23	254	1252	31.7	4.9	87.8%	28.3%	0.4
Village24	163	835	31.9	5.1	93.9%	13.5%	0.5
Village25	252	1313	30.9	5.2	96.4%	30.6%	0.6
Village28	315	1612	31.6	5.1	97.5%	34.0%	0.6
Village29	290	1337	32.2	4.6	84.1%	28.6%	0.6
Village31	153	851	26.1	5.6	97.4%	34.6%	0.5
Village32	241	1181	30.8	4.9	96.7%	20.7%	0.5
Village33	204	843	33.4	4.1	95.1%	6.4%	0.7
Village36	289	1214	33.4	4.2	84.8%	4.8%	0.7
Village39	289	1343	31.8	4.6	93.8%	42.2%	0.7
Village42	192	853	37.7	4.4	89.1%	28.6%	0.7
Village43	198	875	34.1	4.4	97.0%	26.8%	0.7
Village45	222	1076	29.8	4.8	94.6%	34.2%	0.5
Village47	139	687	33.7	4.9	94.2%	38.1%	0.6
Village50	244	999	34.8	4.1	92.2%	26.2%	0.7
Village51	251	1062	33.9	4.2	89.6%	13.1%	0.7
Village52	327	1525	33.8	4.7	91.7%	21.1%	0.7
Village55	257	1180	35.6	4.6	94.9%	4.7%	0.6
Village57	212	956	28.8	4.5	93.9%	3.8%	0.5

Table A.6 Continued: Descriptive Statistics

village	number of households	number of villagers	average age	average family size	household having electric	household having latrine	average rooms per person
Village59	329	1599	31.4	4.9	96.0%	17.9%	0.6
Village62	190	994	32.1	5.2	92.1%	32.6%	0.5
Village65	299	1335	32.8	4.5	93.3%	29.4%	0.7
Village67	193	893	31.8	4.6	96.4%	25.4%	0.6
Village68	153	663	33.0	4.3	88.9%	22.2%	0.7
Village70	205	899	33.1	4.4	95.1%	24.9%	0.7
Village71	298	1388	28.8	4.7	95.0%	42.6%	0.6
Village72	223	999	32.0	4.5	96.9%	30.5%	0.7
Village73	174	870	30.1	5.0	96.6%	20.1%	0.6
Village75	172	831	32.7	4.8	91.3%	27.9%	0.7

Table A.7: Second Stage: who are they

	(1)	(2)	(3)
Agriculture labour	-0.0141	0.0476*	0.0672***
	(0.0136)	(0.0286)	(0.0134)
Anganavadi Teacher	0.0386	0.0664	0.1248**
	(0.0602)	(0.1269)	(0.0593)
Bone Specialist	-0.2170	-0.3465	-0.0213
	(0.3314)	(0.6989)	(0.3265)
Blacksmith	-0.0752	-0.2279	0.1606*
	(0.0927)	(0.1954)	(0.0913)
Construction/mud work	0.0050	0.2199***	0.0562**
	(0.0258)	(0.0544)	(0.0254)
Government Official	-0.0608	-0.0217	0.0350

Table A.7 Continued: Second Stage: who are they

	(1)	(2)	(3)
	(0.0506)	(0.1067)	(0.0498)
Cook	0.0346	0.2015*	-0.0168
	(0.0507)	(0.1068)	(0.0499)
Cow/livestock breeding	0.0059	-0.0235	0.0438
	(0.0282)	(0.0595)	(0.0278)
Truck/Tractor Driver	-0.0401	0.0746	0.0415
	(0.0305)	(0.0642)	(0.0300)
Factory worker (bricks/stones/mill)	0.0175	0.1756***	0.0174
	(0.0246)	(0.0518)	(0.0242)
Milk dairy	0.0595	-0.1104	0.0622
	(0.0931)	(0.1965)	(0.0918)
Poultry farm	-0.1927	0.3577	0.0151
	(0.1258)	(0.2654)	(0.1240)
Small business	0.2006***	0.1287***	0.0606***
	(0.0227)	(0.0479)	(0.0224)
Silk/Cotton work	0.0031	0.0542	0.0266
	(0.0296)	(0.0624)	(0.0292)
Tailor Garment worker	0.0903***	0.1169*	0.0309
	(0.0304)	(0.0642)	(0.0300)
Teacher	0.0268	-0.0452	0.0690
	(0.0426)	(0.0898)	(0.0420)
Daily labourer	-0.0172	0.1239**	0.0390
	(0.0283)	(0.0597)	(0.0279)
Auto driver	0.0113	0.2724**	0.0223
	(0.0548)	(0.1155)	(0.0540)

Table A.7 Continued: Second Stage: who are they

	(1)	(2)	(3)
Police officer	-0.1459	-0.0374	0.3282*
	(0.1917)	(0.4044)	(0.1890)
Waterman	-0.0722	0.0115	0.0715
	(0.0677)	(0.1428)	(0.0667)
Social Worker	-0.1541	-0.2959	-0.0475
	(0.1662)	(0.3505)	(0.1638)
Carpenter	-0.0863	-0.0816	0.0468
	(0.0748)	(0.1578)	(0.0737)
Electronics	0.0711	-0.1140	-0.0337
	(0.0727)	(0.1532)	(0.0716)
Goldsmith	-0.1351	0.2782	-0.0027
	(0.1664)	(0.3510)	(0.1640)
Hotel worker	0.3299***	0.4257***	0.0759
	(0.0750)	(0.1581)	(0.0739)
Poojari	0.3697***	-0.1542	0.1501
	(0.1369)	(0.2887)	(0.1349)
Post man	-0.1708	-0.3427	0.1632
	(0.1253)	(0.2643)	(0.1235)
Veterinary clinic	0.8649***	1.9114***	0.0377
	(0.3314)	(0.6990)	(0.3266)
Mechanic	0.0106	-0.1237	0.1274**
	(0.0634)	(0.1337)	(0.0625)
Painter	-0.0832	0.1570	0.0034
	(0.0746)	(0.1574)	(0.0735)
Real Estate business	0.0158	0.6553***	0.1088

Table A.7 Continued: Second Stage: who are they

	(1)	(2)	(3)
	(0.1108)	(0.2337)	(0.1092)
Skilled labour/work for company	0.0469	0.0252	0.0809*
	(0.0491)	(0.1036)	(0.0484)
Barber/saloon	0.4883***	-0.0036	0.0443
	(0.1005)	(0.2119)	(0.0990)
Lawyer	-0.1235	0.0104	-0.1291
	(0.1915)	(0.4039)	(0.1887)
Security guard	-0.0993	0.0081	0.0016
	(0.1352)	(0.2852)	(0.1332)
Librarian	-0.0451	1.7625**	-0.0848
	(0.3301)	(0.6962)	(0.3253)
Student	-0.2236	0.6929	0.1796
	(0.2341)	(0.4938)	(0.2307)
Doctor/Health assistant	0.2691**	0.2703	0.0874
	(0.1053)	(0.2222)	(0.1038)
Fireman	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)
Photographer	-0.0995	-0.2046	0.2804
	(0.2336)	(0.4926)	(0.2302)
Folk artist	0.3541	-0.4611	0.0144
	(0.2379)	(0.5017)	(0.2344)
Begger	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)
Wood cutter	-0.0223	0.1942	-0.0000
	(0.0600)	(0.1265)	(0.0591)

Table A.7 Continued: Second Stage: who are they

	(1)	(2)	(3)
Musician/Artist	0.3268	0.0640	-0.0436
	(0.2338)	(0.4931)	(0.2304)
Animal skin business	-0.1053	-0.0327	-0.0294
	(0.2350)	(0.4956)	(0.2316)
Average Age	0.0003	-0.0052***	-0.0003
	(0.0006)	(0.0013)	(0.0006)
Electric	0.0234	-0.0204	0.0309
	(0.0229)	(0.0482)	(0.0225)
Latrine	0.0533***	-0.0882***	0.0148
	(0.0134)	(0.0283)	(0.0132)
# Rooms	0.0315***	-0.0087	0.0132***
	(0.0044)	(0.0094)	(0.0044)
Control village fix effect	Y	Y	Y

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

design (1) uses whether one is predefined leaders as response variable

design (2) uses whether one joins the micro-finance program as response variable

design (3) uses whether one is selected by lasso as response variable

APPENDIX B

CHAPTER 2 OF APPENDIX

B.1 Proofs

B.1.1 Proof of Theorem 1

Point

For point discontinuity, the design matrix is exactly the identity matrix I_n and thus the irrepresentable condition holds.

Jump

For general case, let $X_{A_0} = (X_{j_1}, X_{j_2}, \cdots, X_{j_s})$ such that $j_1 > j_2 > \cdots j_s$. Thus

$$X'_{A_0}X_{A_0} = \begin{bmatrix} 1 & \sqrt{\frac{n-j_1}{n-j_2}} & \sqrt{\frac{n-j_1}{n-j_3}} & \cdots & \sqrt{\frac{n-j_1}{n-j_s}} \\ \sqrt{\frac{n-j_1}{n-j_2}} & 1 & \sqrt{\frac{n-j_2}{n-j_3}} & & \vdots \\ \sqrt{\frac{n-j_1}{n-j_3}} & \sqrt{\frac{n-j_2}{n-j_3}} & 1 & & \\ \vdots & & & \ddots & \\ \sqrt{\frac{n-j_1}{n-j_s}} & \cdots & & & 1 \end{bmatrix}$$

write

$$X'_k X_{A_0}(X'_{A_0}X_{A_0})^- = \kappa \Leftrightarrow X'_k X_{A_0} = \kappa(X'_{A_0}X_{A_0})$$

Thus

$$\frac{\min\{n-k, n-j_1\}}{\sqrt{(n-k)(n-j_1)}} = \kappa_1 \sqrt{\frac{n-j_1}{n-j_1}} + \kappa_2 \sqrt{\frac{n-j_1}{n-j_2}} + \kappa_3 \sqrt{\frac{n-j_1}{n-j_3}} + \cdots + \kappa_s \sqrt{\frac{n-j_1}{n-j_s}}$$

$$\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} = \kappa_1 \sqrt{\frac{n-j_1}{n-j_2}} + \kappa_2 \sqrt{\frac{n-j_2}{n-j_2}} + \kappa_3 \sqrt{\frac{n-j_2}{n-j_3}} \cdots + \kappa_s \sqrt{\frac{n-j_2}{n-j_s}}$$

$$\cdots$$

$$\frac{\min\{n-k, n-j_s\}}{\sqrt{(n-k)(n-j_s)}} = \kappa_1 \sqrt{\frac{n-j_1}{n-j_s}} + \kappa_2 \sqrt{\frac{n-j_2}{n-j_s}} + \kappa_3 \sqrt{\frac{n-j_3}{n-j_s}} \cdots + \kappa_s \sqrt{\frac{n-j_s}{n-j_s}}$$

Use equation (1) minus equation (2) times $\sqrt{\frac{n-j_1}{n-j_2}}$

$$\frac{\min\{n-k, n-j_1\}}{\sqrt{(n-k)(n-j_1)}} - \sqrt{\frac{n-j_1}{n-j_2}} \frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} = \kappa_1 \frac{j_1 - j_2}{n-j_2}$$

Thus

$$\kappa_1 = \begin{cases} \sqrt{\frac{n-k}{n-j_1}} & k \geq j_1 \\ \sqrt{\frac{n-j_1}{n-k}} \frac{k-j_2}{j_1-j_2} & j_1 > k \geq j_2 \\ 0 & j_2 > k \end{cases}$$

when $k \geq j_1$, notice that $\kappa_2 = \kappa_3 = \cdots = \kappa_s = 0$ is a solution to the system. And since $(X'_{A_0} X_{A_0})$ is full rank, the solution is unique. Thus

$$\sup_{\|\tau_{A_0}\|_\infty \leq 1} |\kappa \tau_{A_0}| = \sqrt{\frac{n-k}{n-j_1}} < 1$$

when $j_1 > k$, use equation (2) minus equation (3) times $\sqrt{\frac{n-j_2}{n-j_3}}$

$$\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} - \sqrt{\frac{n-j_2}{n-j_3}} \frac{\min\{n-k, n-j_3\}}{\sqrt{(n-k)(n-j_3)}}$$
$$= \kappa_2 \frac{j_2 - j_3}{n-j_3} + \kappa_1 \left(\sqrt{\frac{n-j_1}{n-j_2}} - \frac{\sqrt{(n-j_1)(n-j_2)}}{n-j_3} \right)$$

when $j_1 > k \geq j_2$, substitute κ_1, we have

$$\begin{aligned}
&\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} - \sqrt{\frac{n-j_2}{n-j_3}}\frac{\min\{n-k, n-j_3\}}{\sqrt{(n-k)(n-j_3)}} \\
&\quad = \kappa_2 \frac{j_2 - j_3}{n-j_3} + \kappa_1 \left(\sqrt{\frac{n-j_1}{n-j_2}} - \frac{\sqrt{(n-j_1)(n-j_2)}}{n-j_3} \right)
\end{aligned}$$

Thus,

$$\kappa_2 = \sqrt{\frac{n-j_2}{n-k}}\frac{j_1 - k}{j_1 - j_2}$$

Notice that $\kappa_3 = \kappa_4 = \cdots = \kappa_s = 0$ is a solution to the system.

$$\sup_{\|\tau_{A_0}\|_\infty \leq 1} |\kappa \tau_{A_0}| = \sqrt{\frac{n-j_1}{n-k}}\frac{k-j_2}{j_1-j_2} + \sqrt{\frac{n-j_2}{n-k}}\frac{j_1-k}{j_1-j_2} < 1$$

when $j_2 > k$, since $\kappa_1 = 0$, we can rewrite the system as:

$$\begin{aligned}
\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} &= \kappa_2 \sqrt{\frac{n-j_2}{n-j_2}} + \kappa_3 \sqrt{\frac{n-j_2}{n-j_3}} \cdots + \kappa_s \sqrt{\frac{n-j_2}{n-j_s}} \\
\frac{\min\{n-k, n-j_3\}}{\sqrt{(n-k)(n-j_3)}} &= \kappa_2 \sqrt{\frac{n-j_2}{n-j_3}} + \kappa_3 \sqrt{\frac{n-j_3}{n-j_3}} \cdots + \kappa_s \sqrt{\frac{n-j_3}{n-j_s}} \\
&\cdots \\
\frac{\min\{n-k, n-j_s\}}{\sqrt{(n-k)(n-j_s)}} &= \kappa_2 \sqrt{\frac{n-j_2}{n-j_s}} + \kappa_3 \sqrt{\frac{n-j_3}{n-j_s}} \cdots + \kappa_s \sqrt{\frac{n-j_s}{n-j_s}}
\end{aligned}$$

And we back to the initial system with $s-1$ equations. By induction, we have $\sup_{\|\tau_{A_0}\|_\infty \leq 1} |\kappa \tau_{A_0}|$, thus the irrepresentabile condition holds.

Kink

In the case when only one kink exists, $D'_{A_0} D_{A_0} = 1$ and

$$D_j^{K'} D_{A_0} = \frac{\sum_{i=\max\{k+1, j+1\}}^{n} (x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)})}{\phi_j \phi_k} < 1$$

the irrepresentabile condition holds.

Kink + Jump

For case when there are only one discontinuity but its type (kink or jump) are unknown, the design matrix D is a combine of both jump dummies and kink dummies:

$$D_k^J = \left(0, 0, 0, \cdots, \frac{1}{\sqrt{n-k}}, \frac{1}{\sqrt{n-k}}, \cdots, \frac{1}{\sqrt{n-k}}\right)'$$
$$D_k^K = \left(0, 0, 0, \cdots, \frac{x_{(k+1)} - x_{(k)}}{\phi_k}, \frac{x_{(k+2)} - x_{(k)}}{\phi_k}, \cdots, \frac{x_{(n)} - x_{(k)}}{\phi_k}\right)'$$

where $\phi_k = \sqrt{\sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})^2}$

Since only one break exists, $X'_{A_0} X_{A_0} = 1$

- Assume the break is Jump:

$$X_j^{J'} X_{A_0} = \frac{\min\{(n-k), (n-j)\}}{\sqrt{(n-k)(n-j)}} < 1$$
$$X_j^{K'} X_{A_0} = \frac{\sum_{i=\max\{k+1, j+1\}}^{n} (x_{(i)} - x_{(j)})}{\sqrt{n-j}\phi_j} < 1, \text{(cauchy schwarz)}$$

- Assume the break is kink:

$$X_j^{J'} X_{A_0} = \frac{\sum_{i=\max\{k+1,j+1\}}^{n}(x_{(i)} - x_{(k)})}{\sqrt{n-j}\phi_k} < 1, \text{(cauchy schwarz)}$$
$$X_j^{K'} X_{A_0} = \frac{\sum_{i=\max\{k+1,j+1\}}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)})}{\phi_j \phi_k} < 1$$

B.1.2 Proof of Corollary 1

In the presence of projection, let P_Z be the matrix we project the regressors on. The irrepresentable condition becomes:

$$\begin{aligned}
&\left|X_j' P_Z X_{A_0} (X_{A_0}' P_Z X_{A_0})^{-1} \tau_{A_0}\right| \\
&= \left|trace\left(P_Z X_{A_0} (X_{A_0} P_Z X_{A_0}')^{-1} \tau_{A_0} X_j'\right)\right| \\
&= \left|trace\left((X_{A_0}' X_{A_0})^{-1} X_{A_0} X_{A_0}' P_Z X_{A_0} (X_{A_0} P_Z X_{A_0}')^{-1} \tau_{A_0} X_j'\right)\right| \\
&= \left|trace\left((X_{A_0}' X_{A_0})^{-1} X_{A_0} \tau_{A_0} X_j'\right)\right| \\
&= \left|trace\left(\tau_{A_0} X_j' (X_{A_0}' X_{A_0})^{-1} X_{A_0}\right)\right| \\
&= \left|trace\left(\tau_{A_0} X_j' (X_{A_0}' X_{A_0})^{-1} X_{A_0} X_{A_0}' X_{A_0} (X_{A_0}' X_{A_0})^{-1}\right)\right| \\
&= \left|trace\left(\tau_{A_0} X_j' X_{A_0} (X_{A_0}' X_{A_0})^{-1}\right)\right| \\
&= \left|X_j' X_{A_0} (X_{A_0}' X_{A_0})^{-1} \tau_{A_0}\right| < 1
\end{aligned}$$

□

B.1.3 Proof of Corollary 2

Consider the two kink case at $k_1 > k_2$:

$$X_{A_0}' X_{A_0} = \begin{bmatrix} 1 & \frac{\sum_{i=k_1+1}^{n}(x_{(i)} - x_{(k_1)})(x_{(i)} - x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} \\ \frac{\sum_{i=k_1+1}^{n}(x_{(i)} - x_{(k_1)})(x_{(i)} - x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} & 1 \end{bmatrix}$$

$$(X'_{A_0}X_{A_0})^- = \frac{1}{D}\begin{bmatrix} 1 & -\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} \\ -\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} & 1 \end{bmatrix}$$

where $D = 1 - \left(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}\right)^2$

Now consider $k_1 > j > k_2$

$$X_j^{K'}X_{A_0} = \begin{bmatrix} \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} & \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}} \end{bmatrix}$$

Thus

$$\begin{aligned} & X_j^{K'}X_{A_0}(X'_{A_0}X_{A_0})^-[1,1]' \\ &= \frac{1}{D}\Big(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} \\ &- \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}} \\ &- \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} \\ &+ \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}}\Big) \end{aligned}$$

$$\begin{aligned} &= \frac{1}{D}\Big(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} + \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}}\Big) \\ &\quad \Big(1 - \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}\Big) \\ &= \Big(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} + \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}}\Big) \\ &\quad \Big(1 + \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}\Big)^{-1} \end{aligned}$$

Define $f(j) = \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} + \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}}$

Notice that $f(k_1) = f(k_2) = 1 + \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}$ and f is concave. Thus $X_j^{K'}X_{A_0}(X'_{A_0}X_{A_0})^-[1,1]' > 1$ and the irrepresentabile condition fails. □

B.1.4 Proof of Lemma 1

The problem of (1) is equivalent to the following standard LASSO problem:

$$(\tilde{\psi}) = \arg\min_{\psi} ||M_m Y_n - M_m D_n \psi||_2^2 + \lambda|\psi|_1 \tag{B.1}$$

– where M_m is the projection matrix $I - S_m(S_m' S_m)^{-1} S_m$.

proof. Let $L = ||Y_n - S_m\beta - D_n\psi||_2^2 + \lambda|\psi|_1$

$$\frac{\partial L}{\partial \beta} = -2S_m' Y_n + 2S_m' S_m \beta + 2S_m' D_n \psi \tag{B.2}$$

$\hat{\beta}$ must satisfies $\frac{\partial L}{\partial \beta} = 0$. Thus,

$$\hat{\beta} = (S_m' S_m)^{-1} S_m'(Y_n - D_n\psi) \tag{B.3}$$

Substitute (8) into (2), we have:

$$\begin{aligned} L &= ||(I - P_S)(Y_n - D_n\psi)||_2^2 + \lambda|\psi|_1 \\ &= ||M_m Y_n - M_m D_n \psi||_2^2 + \lambda|\psi|_1 \end{aligned} \tag{B.4}$$

□

B.1.5 Proof of Theorem 2

Let ω_m be the compatibility constant for the sequence of projection $I - S_m(S_m' S_m)^{-1} S_m'$. If there exists ω such that for all $m > m_0$, $|\omega_m| > \omega$

$$IMSE_n(m) \leq 2\phi_m^2 + 2\sigma^2 \frac{K_m}{n} + 8W_1^2\sigma^2 \frac{\log p_n}{n} s_0/\omega^2 \tag{B.5}$$

proof.

Theorem 1 and 2 show that irrepresentable condition holds for our design matrix D as well as a projection of it $P_S D$ for any projection matrix P_S. From Van de geer book, we know that irrepresentable condition implies compatibility condition.

Compatibility Condition.

We say that the compatibility condition is met for the set A_0 if for some $\omega_m > 0$ (independent of n) and for all ψ satisfying $\|\psi_{A_0^c}\|_1 \leq 3\|\psi_{A_0}\|_1$, it holds that

$$\|\psi_{A_0}\|_1^2 \leq (\psi' \hat{\Sigma}_m \psi) s_{A_0} / \omega_m^2, \tag{B.6}$$

. where s_{A_0} is the number of elements in A_0. $\hat{\Sigma}_m = (M_m D_n)' M_m D_n$.

We show that in theorem 1 that irrepresentable condition holds for our design matrix $\hat{\Sigma}_m = (M_m D_n)' M_m D_n$. irrepresentable condition implies compatibility condition as shown in Van de geer book.

Lower bound for compatibility constant.

Let $\{M_m\}_m$ be a set of projection matrix. And Let $\{\omega_m^2\}_m$ be the set of compatibility constant associated with design $\hat{\Sigma}_m = (M_m D_n)' M_m D_n$. There exist a constant ω^2 such that $\omega_m^2 > \omega^2$ for all $m > m_0$.

The compatibility assumptions leads to the following results: when $\lambda =$

$W_1\sigma\sqrt{\frac{\log p}{n}}$ for some constant W_1,

$$\|M_m D_n(\hat{\psi} - \psi)\|_2^2/n \leq 4\lambda^2 s_0/\omega_m^2 \tag{B.7}$$

$$\|(\hat{\psi} - \psi)\|_2^2 \leq \|(\hat{\psi} - \psi)\|_1^2 \leq 16\lambda^2 s_0^2/\omega_m^4 \tag{B.8}$$

Assume

$$y = g(x) + \sum_{k=1}^{K} d_k(x)\psi_k + \epsilon.$$

further, define

$$r_m(x) = g(x) - S_m(x)'\beta_m$$

Let $(\hat{\beta}, \hat{\psi})$ be estimates from the lasso. Notice that

$$\begin{aligned}
\hat{\beta} &= (S_m' S_m)^{-1} S_m'(Y_n - D_n\psi) = (S_m' S_m)^{-1} S_m' \left(g(X) + \sum_{k=1}^{K} d_k(X)\psi_k + \epsilon - D_n\hat{\psi} \right) \\
&= (S_m' S_m)^{-1} S_m' \left(g(X) + \epsilon \right) + (S_m' S_m)^{-1} S_m' \left(\sum_{k=1}^{K} d_k(X)\psi_k - D_n\hat{\psi} \right) \\
&= \tilde{\beta}_m - (S_m' S_m)^{-1} S_m' \left(D_n(\hat{\psi} - \psi) \right)
\end{aligned}$$

where $\tilde{\beta} = (S_m' S_m)^{-1} S_m' \left(g(X) + \epsilon \right)$ are coefficients for a standard SEIVE regression on $g(X) + \epsilon$.

$$
\begin{aligned}
IMSE_n(m) &= \int \mathbb{E}\left(\hat{g}_m(x) - g(x) + D_n(x)'\hat{\psi} - \sum_{k=1}^{K} d_k(x)\psi_k\right)^2 f(x)dx \\
&= \int \mathbb{E}\left(\hat{g}_m(x) - g(x) + D_n(x)'(\hat{\psi} - \psi)\right)^2 f(x)dx \\
&= \int \mathbb{E}\left(S_m(x)'(\tilde{\beta}_m - \beta_m) + r_m(x) - S_m(x)(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi) + D(x)'(\hat{\psi} - \psi)\right)^2 f(x)dx \\
&\leq 2\underbrace{\int \mathbb{E}(S_m(x)(\tilde{\beta}_m - \beta_m) + r_m(x))^2 f(x)dx}_{(A1)} \\
&+ 2\underbrace{\int \mathbb{E}\left(D_n(x)'(\hat{\psi} - \psi) - S_m(x)(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi)\right)^2 f(x)dx}_{(A2)}
\end{aligned}
$$

Notice that part (A1) is the standard SEIVE term and as shown in Hansen 2014,

$$
(A1) = \phi_m^2 + \sigma^2\frac{K_m}{n}
$$

How consider part (A2). Define $\mathbb{E}(S_m(x)'S_m(x)) = Q_m$. Define

$$
DD_n = \int D_n(x)D_n(x)'f(x)dx = \begin{bmatrix}
\frac{\mathbb{P}(x>x_{(1)})}{n-1} & \frac{\mathbb{P}(x>x_{(2)})}{\sqrt{n-1}\sqrt{n-2}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-1}\sqrt{1}} \\
\frac{\mathbb{P}(x>x_{(2)})}{\sqrt{n-1}\sqrt{n-2}} & \frac{\mathbb{P}(x>x_{(2)})}{n-2} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-2}\sqrt{1}} \\
\vdots & & & \vdots \\
\frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-1}\sqrt{1}} & \frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-2}\sqrt{1}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})}{1}
\end{bmatrix}
$$

Notice that

$$D_n' D_n = \begin{bmatrix} \frac{\mathbf{1}(x_1>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_2>x_{(1)})}{\sqrt{n-1}} & \cdots & \frac{\mathbf{1}(x_n>x_{(1)})}{\sqrt{n-1}} \\ \frac{\mathbf{1}(x_1>x_{(2)})}{\sqrt{n-2}} & \frac{\mathbf{1}(x_2>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_n>x_{(2)})}{\sqrt{n-2}} \\ \vdots & & & \vdots \\ \frac{\mathbf{1}(x_1>x_{(n-1)})}{\sqrt{1}} & \frac{\mathbf{1}(x_2>x_{(n-1)})}{\sqrt{1}} & \cdots & \frac{\mathbf{1}(x_n>x_{(n-1)})}{\sqrt{1}} \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}(x_1>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_1>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_1>x_{(n-1)})}{\sqrt{1}} \\ \frac{\mathbf{1}(x_2>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_2>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_2>x_{(n-1)})}{\sqrt{1}} \\ \vdots & & & \vdots \\ \frac{\mathbf{1}(x_n>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_n>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_n>x_{(n-1)})}{\sqrt{1}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \sqrt{\frac{n-2}{n-1}} & \cdots & \sqrt{\frac{1}{n-1}} \\ \sqrt{\frac{n-2}{n-1}} & 1 & \cdots & \sqrt{\frac{1}{n-2}} \\ \vdots & & & \vdots \\ \sqrt{\frac{1}{n-1}} & \sqrt{\frac{1}{n-2}} & \cdots & 1 \end{bmatrix}$$

and

$$DD_n - \frac{1}{n} D_n' D_n = \begin{bmatrix} \frac{\mathbb{P}(x>x_{(1)}) - \frac{n-1}{n}}{n-1} & \frac{\mathbb{P}(x>x_{(2)}) - \frac{n-2}{n}}{\sqrt{n-1}\sqrt{n-2}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)}) - \frac{1}{n}}{\sqrt{n-1}\sqrt{1}} \\ \frac{\mathbb{P}(x>x_{(2)}) - \frac{n-2}{n}}{\sqrt{n-1}\sqrt{n-2}} & \frac{\mathbb{P}(x>x_{(2)}) - \frac{n-2}{n}}{n-2} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)}) - \frac{1}{n}}{\sqrt{n-2}\sqrt{1}} \\ \vdots & & & \vdots \\ \frac{\mathbb{P}(x>x_{(n-1)}) - \frac{1}{n}}{\sqrt{n-1}\sqrt{1}} & \frac{\mathbb{P}(x>x_{(n-1)}) - \frac{1}{n}}{\sqrt{n-2}\sqrt{1}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)}) - \frac{1}{n}}{1} \end{bmatrix}$$

Let $\hat{\mathbb{P}}$ be the empirical distribution. Since $\hat{\mathbb{P}}(x > x_{(i)}) = \frac{n-i}{n}$, thus

$$\|DD_n - \frac{1}{n} D_n' D_n\|_\infty \to 0 \quad a.s.$$

Second, Define

$$DS_n = \int D_n(x) S_m(x)' f(x) dx = \begin{bmatrix} \frac{\int \mathbf{1}(x>x_{(1)}) S_m(x)' f(x) dx}{\sqrt{n-1}} \\ \frac{\int \mathbf{1}(x>x_{(2)}) S_m(x)' f(x) dx}{\sqrt{n-2}} \\ \vdots \\ \frac{\int \mathbf{1}(x>x_{(n-1)}) S_m(x)' f(x) dx}{\sqrt{1}} \end{bmatrix} = \begin{bmatrix} \frac{\mathbb{E}(S_m(x)' | x>x_{(1)})}{\sqrt{n-1}} \\ \frac{\mathbb{E}(S_m(x)' | x>x_{(2)})}{\sqrt{n-2}} \\ \vdots \\ \frac{\mathbb{E}(S_m(x)' | x>x_{(n-1)})}{\sqrt{1}} \end{bmatrix}$$

$$D_n' S_m = \begin{bmatrix} \frac{\mathbf{1}(x_1 > x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_2 > x_{(1)})}{\sqrt{n-1}} & \cdots & \frac{\mathbf{1}(x_n > x_{(1)})}{\sqrt{n-1}} \\ \frac{\mathbf{1}(x_1 > x_{(2)})}{\sqrt{n-2}} & \frac{\mathbf{1}(x_2 > x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_n > x_{(2)})}{\sqrt{n-2}} \\ \vdots & & & \vdots \\ \frac{\mathbf{1}(x_1 > x_{(n-1)})}{\sqrt{1}} & \frac{\mathbf{1}(x_2 > x_{(n-1)})}{\sqrt{1}} & \cdots & \frac{\mathbf{1}(x_n > x_{(n-1)})}{\sqrt{1}} \end{bmatrix} \begin{bmatrix} S_m(x_1)' \\ S_m(x_2)' \\ \vdots \\ S_m(x_n)' \end{bmatrix}$$

$$= \int D_n(x) S_m(x)' f(x) dx = \begin{bmatrix} \frac{\sum_{i=1}^n \mathbf{1}(x_i > x_{(1)}) S_m(x_i)'}{\sqrt{n-1}} \\ \frac{\sum_{i=1}^n \mathbf{1}(x_i > x_{(2)}) S_m(x_i)'}{\sqrt{n-2}} \\ \vdots \\ \frac{\sum_{i=1}^n \mathbf{1}(x_i > x_{(n-1)}) S_m(x_i)'}{\sqrt{1}} \end{bmatrix}$$

$$(A2) = \int \left(D_n(x)'(\hat{\psi} - \psi) - S_m(x)'(S_m' S_m)^{-1} S_m' D_n(\hat{\psi} - \psi) \right)^2 f(x) dx$$

$$= (\hat{\psi} - \psi)' \underbrace{\left(\int (D_n(x)' - S_m(x)'(S_m' S_m)^{-1} S_m' D_n)' (D_n(x)' - S_m(x)'(S_m' S_m)^{-1} S_m' D_n) f(x) dx \right)}_{(A3)} (\hat{\psi} - \psi)$$

$$(A3) = \left(\int (D_n(x)' - S_m(x)'(S_m' S_m)^{-1} S_m' D_n)' (D_n(x)' - S_m(x)'(S_m' S_m)^{-1} S_m' D_n) f(x) dx \right)$$

$$= \Big(\int (D_n(x) D_n(x)' - D_n(x) S_m(x)'(S_m' S_m)^{-1} S_m' D_n$$

$$- D_n S_m (S_m' S_m)^{-1} S_m(x) D_n(x)' + D_n' S_m (S_m' S_m)^{-1} S_m(x) S_m(x)' (S_m' S_m)^{-1} S_m' D_n) f(x) dx \Big)$$

$$= \Big(\int D_n(x) D_n(x)' f(x) dx \Big) - \Big(\int D_n(x) S_m(x)' f(x) dx (S_m' S_m)^{-1} S_m' D_n \Big)$$

$$- \Big(D_n' S_m (S_m' S_m)^{-1} \int S_m(x) D_n(x)' f(x) dx \Big) + \Big(D_n' S_m (S_m' S_m)^{-1} \int S_m(x) S_m(x)' f(x) dx (S_m' S_m)^{-1} S_m' D_n \Big)$$

$$= DD_n - DS_n (S_m' S_m)^{-1} S_m' D_n - D_n' S_m (S_m' S_m)^{-1} (DS_n') + \Big(D_n' S_m (S_m' S_m)^{-1} Q_m (S_m' S_m)^{-1} S_m' D_n \Big)$$

$$= \underbrace{\frac{1}{n} D_n' (I - S_m (S_m' S_m)^{-1} S_m') D_n}_{(A4)} + \underbrace{\Big(DD_n - \frac{1}{n} D_n' D_n \Big)}_{(A5)} - \underbrace{\Big(DS_n - \frac{1}{n} D_n S_m \Big)(S_m' S_m)^{-1} S_m' D_n}_{(A6)}$$

$$- \underbrace{D_n' S_m (S_m' S_m)^{-1} \Big(DS_n - \frac{1}{n} D_n S_m \Big)'}_{(A7)} + \underbrace{\Big(D_n' S_m (S_m' S_m)^{-1} \Big(Q_m - \frac{1}{n} S_m' S_m \Big)(S_m' S_m)^{-1} S_m' D_n \Big)}_{(A8)}$$

For part (A4),

$$(\hat{\psi}-\psi)'\frac{1}{n}D_n'(I-S_m(S_m'S_m)^{-1}S_m')D_n(\hat{\psi}-\psi)=\|M_mD_n(\hat{\psi}-\psi)\|_2^2/n\le 4\lambda^2 s_0/\omega_m^2\le 4\lambda^2 s_0/\omega^2$$

For part (A5),

$$(\hat{\psi}-\psi)'\Big(DD_n-\frac{1}{n}D_n'D_n\Big)(\hat{\psi}-\psi)\le o(1/n)\|(\hat{\psi}-\psi)\|_2^2\le o(1/n)16\lambda^2 s_0^2/\omega_m^4\le o(1/n)16\lambda^2 s_0^2/\omega^4$$

Similarly, (A6), (A7) and (A8) can be show as order $o(1/n)16\lambda^2 s_0^2/\omega^4$.

As a result,

$$IMSE_n(m)\le 2\phi_m^2+2\sigma^2\frac{K_m}{n}+8W_1^2\sigma^2\frac{\log p_n}{n}s_0/\omega^2$$

□

B.1.6 proof of theorem 3

Let $\{(y_i,x_i)\}_{i=1}^n$ be observed data. Assume $y_i=z_i^0\tilde{\beta}+\epsilon_i$, such that

$$z_i^0=\begin{cases}0 & \text{if } x_i\le\gamma\\ 1 & \text{if } x_i>\gamma\end{cases}\tag{B.9}$$

Let $\{(x_{(1)},x_{(2)},\cdots,x_{(n)})\}$ be the ordered statistic of $\{x_i\}_{i=1}^n$. Let $k>0$ be an integer such that

$$x_{(k)}\le\gamma<x_{(k+1)}$$

Now define

$$z_i^{(j)} = \begin{cases} 0 & \text{if } x_i \leq x_{(j)} \\ \frac{1}{\sqrt{n-j}} & \text{if } x_i > x_{(j)} \end{cases} \tag{B.10}$$

We can rewrite $y_i = z_i^{(k)}\beta + \epsilon_i$, where $\beta = \sqrt{n-k}\tilde{\beta}$. Thus:

$$\begin{aligned} z^{(j)'}Y &= \sum_{i=1}^{n} z_i^{(j)} z_i^0 \beta + z_i^{(j)}\epsilon \\ &= \begin{cases} \frac{(n-k)\beta}{\sqrt{n-j}} + \frac{\sum_{i=j+1}^{n}\epsilon_i}{\sqrt{n-j}} & \text{if } j \leq k \\ \\ \frac{(n-j)\beta}{\sqrt{n-j}} + \frac{\sum_{i=j+1}^{n}\epsilon_i}{\sqrt{n-j}} & \text{if } j > k \end{cases} \end{aligned} \tag{B.11}$$

Define $V(j) = (z^{(j)'}Y)^2$, and $\hat{k} = \arg\max_j V(j)$

$$\begin{aligned} V(j) - V(k) &= (z^{(j)'}Y)^2 - (z^{(k)'}Y)^2 \\ &= Y'(z^{(j)}z^{(j)'} - z^{(k)}z^{(k)'})Y \\ &= \beta z^{(k)'}\left(z^{(j)}z^{(j)'} - z^{(k)}z^{(k)'}\right)z^{(k)}\beta + h(\epsilon,\beta) \\ &= \beta\left(z^{(k)'}z^{(j)}z^{(j)'}z^{(k)} - 1\right)\beta + h(\epsilon,\beta) \end{aligned}$$

$$h(\epsilon,\beta) = 2\beta z^{(k)'}(z^{(j)}z^{(j)'} - z^{(k)}z^{(k)'})\epsilon + \epsilon'(z^{(j)}z^{(j)'} - z^{(k)}z^{(k)'})\epsilon$$

Define

$$\gamma(j) = \frac{\beta\left(1 - z^{(k)'}z^{(j)}z^{(j)'}z^{(k)}\right)\beta}{|j-k|}$$

if $j > k$

$$\gamma(j) = \frac{\beta^2}{n-k} = \tilde{\beta}^2$$

if $j < k$

$$\gamma(j) = \frac{\beta^2}{n-j} = \frac{n-k}{n-j}\tilde{\beta}^2$$

Now consistency of $\hat{k}$: $P(|\hat{k}-k|>C)<\xi$:

$$\begin{aligned}
P(|\hat{k}-k|>C) &= P(\sup_{|j-k|>C} V(j) \geq V(k)) \\
&\leq P(\sup_{|j-k|>C} |h(\epsilon,\beta)| \geq \inf_{|j-k|>C} |j-k|\gamma(j)) \\
&\leq P\left(\sup_{|j-k|>C} \left|\frac{h(\epsilon,\beta)}{j-k}\right| \geq \inf_{|j-k|>C} \gamma(j)\right) \\
&\leq P\left(\sup_{|j-k|>C} \left|\frac{h(\epsilon,\beta)}{j-k}\right| \geq \gamma_K\right)
\end{aligned}$$

where $\gamma_K = \tilde{\beta}^2 > 0$

when $j > k$

$$\begin{aligned}
\epsilon'(z^{(j)}z^{(j)'})\epsilon - \epsilon'(z^{(k)}z^{(k)'})\epsilon &= \frac{1}{n-j}\Big(\sum_{l=j+1}^{n}\epsilon_l\Big)^2 - \frac{1}{n-k}\Big(\sum_{l=k+1}^{n}\epsilon_l\Big)^2 \\
&= \frac{j-k}{(n-j)(n-k)}\Big(\sum_{l=j+1}^{n}\epsilon_l\Big)^2 - \frac{2}{n-k}\Big(\sum_{l=k+1}^{j}\epsilon_l\Big)\Big(\sum_{l=j+1}^{n}\epsilon_l\Big) \\
&\quad - \frac{1}{n-k}\Big(\sum_{l=k+1}^{j}\epsilon_l\Big)^2 \\
&= (j-k)o(1)
\end{aligned}$$

similarly, one can show the same rate of convergence when $j < k$.

Now consider $2\beta z^{(k)'}(z^{(j)}z^{(j)'} - z^{(k)}z^{(k)'})\epsilon$.

When $k > j$

$$\begin{aligned} 2\beta z^{(k)'} z^{(j)} z^{(j)'} \epsilon - 2\beta z^{(k)'} z^{(k)} z^{(k)'} \epsilon &= 2\tilde{\beta} \frac{n-k}{n-j} \sum_{l=j+1}^{n} \epsilon_l - 2\tilde{\beta} \sum_{l=k+1}^{n} \epsilon_l \\ &= 2\tilde{\beta} \frac{j-k}{n-j} \sum_{l=k+1}^{n} \epsilon_l + 2\tilde{\beta} \frac{n-k}{n-j} \sum_{l=j+1}^{k} \epsilon_l \\ &= (j-k) o(1) \end{aligned}$$

and for any ξ, we can find a C such that $P(|\hat{k} - k| > C) < \xi$.

For Asymptotic, consider $\tilde{\beta}_n \to 0$ but $\sqrt{n}\tilde{\beta}_n \to \infty$. In this case, $\gamma(j)$ can no longer be treated as $O_p(1)$ but $O_p(\tilde{\beta}_n^2)$. Thus, the convergence of $\hat{k}$ is

$$\hat{k} = k + O_p(\|\tilde{\beta}_n\|^{-2})$$

Let

$$\hat{k} = k + v\lambda_n^{-2}$$

where $\lambda = O_p(\|\tilde{\beta}_n\|)$ and v is a real number in a compact set. Let

$$K(B) = \{w : w = k + [v\lambda_n^{-2}], |v| < B\}$$

we derive the limiting process of $V(w) - V(k)$ for $w \in K(B)$ and then use continuous mapping theorem for arg max to derive the asymptotic distribution. Recall that $\tilde{\beta} = \beta / \sqrt{n-k}$

$$V(w) - V(k) = \beta\Big(z^{(k)'} z^{(w)} z^{(w)'} z^{(k)} - 1\Big)\beta + 2\beta z^{(k)'} (z^{(w)} z^{(w)'} - z^{(k)} z^{(k)'})\epsilon + \epsilon'(z^{(w)} z^{(w)'} - z^{(k)} z^{(k)'})\epsilon$$

when $w > k$,

$$z^{(k)'} z^{(w)} z^{(w)'} z^{(k)} = \frac{n-w}{n-k}$$

Thus

$$\beta\Big(z^{(k)'} z^{(w)} z^{(w)'} z^{(k)} - 1\Big)\beta = -\beta^2\Big(\frac{w-k}{n-k}\Big) = -\beta^2\Big(\frac{[v\lambda_n^{-2}]}{n-k}\Big) = -v$$

next,

$$z^{(k)'}(z^{(w)} z^{(w)'})\epsilon = \frac{\sum_{l=w+1}^{n} \epsilon_l}{\sqrt{n-k}}$$

$$z^{(k)'}(z^{(k)} z^{(k)'})\epsilon = \frac{\sum_{l=k+1}^{n} \epsilon_l}{\sqrt{n-k}}$$

So

$$\beta z^{(k)'}(z^{(w)} z^{(w)'})\epsilon - \beta z^{(k)'}(z^{(k)} z^{(k)'})\epsilon = -\tilde{\beta} \sum_{l=k+1}^{w} \epsilon_l = -\tilde{\beta} \frac{v\lambda_n^{-2}}{w-k} \sum_{l=k+1}^{w} \epsilon_l$$

$$\rightarrow_d -W_1(v)$$

where W_1 is a wiener process of degree v.

And finally,

$$\begin{aligned}
\epsilon'(z^{(w)} z^{(w)'} - z^{(k)} z^{(k)'})\epsilon &= \frac{1}{n-w}\Big(\sum_{l=w+1}^{n} \epsilon_l\Big)^2 - \frac{1}{n-k}\Big(\sum_{l=k+1}^{n} \epsilon_l\Big)^2 \\
&= \frac{w-k}{(n-w)(n-k)}\Big(\sum_{l=k+1}^{n} \epsilon_l\Big)^2 - \frac{1}{n-k}\Big(\sum_{l=k+1}^{w} \epsilon_l\Big)^2 - \frac{2}{n-k}\Big(\sum_{l=k+1}^{w} \epsilon_l\Big)\Big(\sum_{l=k+1}^{n} \epsilon_l\Big) \\
&= o_p(1)
\end{aligned}$$

When $w < k$

$$z^{(k)'} z^{(w)} z^{(w)'} z^{(k)} = \frac{n-k}{n-w}$$

Thus

$$\beta\Big(z^{(k)'} z^{(w)} z^{(w)'} z^{(k)} - 1\Big)\beta = -\beta^2\Big(\frac{k-w}{n-w}\Big) = \beta^2\Big(\frac{[v\lambda_n^{-2}]}{n-w}\Big) = \tilde{\beta}^2 v\lambda_n^{-2}$$

next,

$$z^{(k)'}(z^{(w)} z^{(w)'})\epsilon = \frac{\sqrt{n-k}}{n-w}\sum_{l=w+1}^{n}\epsilon_l$$

$$z^{(k)'}(z^{(k)} z^{(k)'})\epsilon = \frac{\sum_{l=k+1}^{n}\epsilon_l}{\sqrt{n-k}}$$

So

$$\begin{aligned}\beta z^{(k)'}(z^{(w)} z^{(w)'})\epsilon - \beta z^{(k)'}(z^{(k)} z^{(k)'})\epsilon &= \beta\frac{\sqrt{n-k}}{n-w}\sum_{l=w+1}^{k}\epsilon_l + \beta\frac{w-k}{n-w}\frac{\sum_{l=k+1}^{n}\epsilon_l}{\sqrt{n-k}} \\ &= \tilde{\beta}\frac{n-k}{n-w}\sum_{l=w+1}^{k}\epsilon_l + \sqrt{n-k}\tilde{\beta}\frac{w-k}{n-w}\frac{\sum_{l=k+1}^{n}\epsilon_l}{\sqrt{n-k}} \\ &\to_d W_2(v)\end{aligned}$$

where W_2 is another wiener process of degree v.

Thus

$$V(k + [v\lambda_n^{-2}]) - V(k) \to_d -|v| + 2W(|v|)$$

By continuous mapping theorem:

$$\lambda_n^2(\hat{k} - k) \to_d \arg\max_v(-|v| + 2W(|v|)) \tag{B.12}$$

B.1.7 proof of theorem 4

Let $\{(y_i, x_i)\}_{i=1}^n$ be observed data. Assume $y_i = u_i^0\tilde{\beta} + \epsilon_i$, such that

$$u_i^0 = \begin{cases} 0 & \text{if } x_i \le z \\ (x_i - z) & \text{if } x_i > z \end{cases} \tag{B.13}$$

Let $\{(x_{(1)}, x_{(2)}, \cdots, x_{(n)})\}$ be the ordered statistic of $\{x_i\}_{i=1}^n$. Let $k > 0$ be an integer such that

$$x_{(k)} \le z < x_{(k+1)}$$

Now define

$$u_i^{(j)} = \begin{cases} 0 & \text{if } x_i \le x_{(j)} \\ \frac{(x_i - x_{(j)})}{\phi_j} & \text{if } x_i > x_{(j)} \end{cases}$$

where $\phi_j = \sqrt{\sum_{i=j+1}^n (x_{(i)} - x_{(j)})^2}$

We can rewrite $y_i = (x_i - z)_+\tilde{\beta} + \epsilon_i$. Thus:

$$\begin{aligned} u^{(j)'}Y &= \sum_{i=1}^n u_i^{(j)'}(x_i - z)_+\tilde{\beta} + u_i^{(j)'}\epsilon \\ &= \begin{cases} \frac{\sum_{i=k+1}^n (x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta} + \frac{\sum_{i=j+1}^n (x_{(i)}-x_{(j)})\epsilon_i}{\phi_j} & \text{if } j \le k \\ \frac{\sum_{i=j+1}^n (x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta} + \frac{\sum_{i=j+1}^n (x_{(i)}-x_{(j)})\epsilon_i}{\phi_j} & \text{if } j > k \end{cases} \end{aligned}$$

Define $V(j) = (u^{(j)'}Y)^2$, and $\hat{k} = \arg\max_j V(j)$.

$$\begin{aligned} V(j) - V(k) &= (u^{(j)'}Y)^2 - (u^{(k)'}Y)^2 \\ &= \begin{cases} \left(\frac{\sum_{i=k+1}^n (x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\right)^2 - \left(\frac{\sum_{i=k+1}^n (x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\right)^2 + h_1(\epsilon,\beta) & \text{if } j \le k \\ \left(\frac{\sum_{i=j+1}^n (x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\right)^2 - \left(\frac{\sum_{i=k+1}^n (x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\right)^2 + h_2(\epsilon,\beta) & \text{if } j \le k \end{cases} \end{aligned}$$

$$h_1(\epsilon,\beta) = \Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)^2 + 2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)$$
$$- \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)^2 + 2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)$$

$$h_2(\epsilon,\beta) = \Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)^2 + 2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)$$
$$- \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)^2 + 2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)$$

When $j \le k$

$$\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$= \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)}) + (x_{(k)}-z)\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2$$
$$- \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)}) + (x_{(k)}-z)\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$= \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$+ \Big((x_{(k)}-z)\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)$$
$$- \Big((x_{(k)}-z)\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_k^2}\tilde{\beta}^2\Big)$$
$$+ \Big((x_{(k)}-z)\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \Big((x_{(k)}-z)\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$= \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2 + O((x_{(k)}-z))$$

Since $(x_{(k)}-z)$ is of order $(1/n)$, we can focus on the first term.

$$\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$= \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)}) + \sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(k)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \phi_k^2\tilde{\beta}^2$$
$$= \frac{\phi_k^4}{\phi_j^2}\tilde{\beta}^2 + 2\Big((x_{(k)}-x_{(j)})\frac{\phi_k^2\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_j^2}\tilde{\beta}^2\Big) + \Big((x_{(k)}-x_{(j)})\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_j}\tilde{\beta}\Big)^2 - \phi_k^2\tilde{\beta}^2$$

Consider the numerator:

$$\phi_k^4\tilde{\beta}^2 + 2\Big((x_{(k)} - x_{(j)})\phi_k^2 \sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\tilde{\beta}^2\Big) + \Big((x_{(k)} - x_{(j)}) \sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\tilde{\beta}\Big)^2 - \phi_j^2\phi_k^2\tilde{\beta}^2$$

$$\begin{aligned}
\phi_j^2\phi_k^2 &= \phi_k^2 \sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})^2 = \phi_k^2 \sum_{i=k+1}^{n}(x_{(i)} - x_{(j)})^2 + \phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2 \\
&= \phi_k^2\Big(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)} + x_{(k)} - x_{(j)})^2\Big) + \phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2 \\
&= \phi_k^2\Big(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})^2 + (n-k)(x_{(k)} - x_{(j)})^2 + 2(x_{(k)} - x_{(j)}) \sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\Big) \\
&+ \phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2 \\
&= \phi_k^4 + 2\phi_k^2(x_{(k)} - x_{(j)}) \sum_{i=k+1}^{n}(x_{(i)} - x_{(k)}) + \phi_k^2(n-k)(x_{(k)} - x_{(j)})^2 + \phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2
\end{aligned}$$

Since

$$\begin{aligned}
&\phi_k^2(n-k)(x_{(k)} - x_{(j)})^2 - \Big((x_{(k)} - x_{(j)}) \sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\Big)^2 \\
&= (x_{(k)} - x_{(j)})^2\Big(\phi_k^2(n-k) - \Big(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\Big)^2\Big) > 0 \quad (Cauchy)
\end{aligned}$$

Notice that $(x_{(k)} - x_{(j)}) = O(1/n)$, thus when $j < k$,

$$\begin{aligned}
&\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)} - z)(x_{(i)} - x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)} - z)(x_{(i)} - x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2 \\
&= -\tilde{\beta}^2(x_{(k)} - x_{(j)})^2\Big(\phi_k^2(n-k) - \Big(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\Big)^2\Big)/\phi_j^2 - \tilde{\beta}^2\phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2/\phi_j^2 \\
&= -\tilde{\beta}^2\phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2/\phi_j^2
\end{aligned}$$

When $j > k$

$$\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$=\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})+(x_{(k)}-z)\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2$$
$$-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})+(x_{(k)}-z)\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$=\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$+\Big((x_{(k)}-z)\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)$$
$$-\Big((x_{(k)}-z)\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_k^2}\tilde{\beta}^2\Big)$$
$$+\Big((x_{(k)}-z)\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-\Big((x_{(k)}-z)\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$
$$=\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2+O((x_{(k)}-z))$$

Notice that

$$\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2$$

$$=\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})+\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(k)}-x_{(j)})-\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2$$

$$-\phi_k^2\tilde{\beta}^2$$

$$=\frac{\phi_k^4}{\phi_j^2}\tilde{\beta}^2+2\Big((x_{(k)}-x_{(j)})\frac{\phi_k^2\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_j^2}\tilde{\beta}^2\Big)+\Big((x_{(k)}-x_{(j)})\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})}{\phi_j}\tilde{\beta}\Big)^2-\phi_k^2\tilde{\beta}^2$$

$$+\Big(\frac{\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2$$

$$-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)$$

$$=\frac{\phi_k^4}{\phi_j^2}\tilde{\beta}^2+\Big(\frac{\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2$$

$$-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)$$

$$+2(x_{(k)}-x_{(j)})\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)-\phi_k^2\tilde{\beta}^2+O(1/n)$$

$$=\frac{\phi_k^4}{\phi_j^2}\tilde{\beta}^2+\Big(\frac{\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-2\Big(\frac{\phi_k^2\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)-\phi_k^2\tilde{\beta}^2+O(1/n)$$

$$=\frac{\phi_k^4}{\phi_j^2}\tilde{\beta}^2+\Big(\frac{\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-2\Big(\frac{\phi_k^2\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)$$

$$+2(x_{(j)}-x_{(k)})\Big(\frac{\phi_k^2\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})}{\phi_j^2}\tilde{\beta}^2\Big)-\phi_k^2\tilde{\beta}^2+O(1/n)$$

$$=\frac{\phi_k^4}{\phi_j^2}\tilde{\beta}^2+\Big(\frac{\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2-2\Big(\frac{\phi_k^2\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})^2}{\phi_j^2}\tilde{\beta}^2\Big)-\phi_k^2\tilde{\beta}^2+O(1/n)$$

Also,

$$\begin{aligned}
\phi_j^2\phi_k^2 &= \phi_k^2 \sum_{i=j+1}^{n} (x_{(i)} - x_{(j)})^2 = \phi_k^2 \sum_{i=k+1}^{n} (x_{(i)} - x_{(j)})^2 - \phi_k^2 \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2 \\
&= \phi_k^2\Big(\sum_{i=k+1}^{n} (x_{(i)} - x_{(k)} + x_{(k)} - x_{(j)})^2\Big) - \phi_k^2 \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2 \\
&= \phi_k^2\Big(\sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})^2 + (n-k)(x_{(k)} - x_{(j)})^2 + 2(x_{(k)} - x_{(j)}) \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})\Big) \\
&- \phi_k^2 \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2 \\
&= \phi_k^4 + 2\phi_k^2(x_{(k)} - x_{(j)}) \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)}) + \phi_k^2(n-k)(x_{(k)} - x_{(j)})^2 - \phi_k^2 \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2 \\
&= \phi_k^4 - \phi_k^2 \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2 + O(n)
\end{aligned}$$

thus when $j > k$,

$$\begin{aligned}
&\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)} - z)(x_{(i)} - x_{(j)})}{\phi_j}\tilde{\beta}\Big)^2 - \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)} - z)(x_{(i)} - x_{(k)})}{\phi_k}\tilde{\beta}\Big)^2 \\
&= -\tilde{\beta}^2\phi_k^2 \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2/\phi_j^2
\end{aligned}$$

Define:

$$\gamma(j) = \begin{cases} \tilde{\beta}^2\phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2/\phi_j^2 & j \le k \\ \tilde{\beta}^2\phi_k^2 \sum_{i=k+1}^{j}(x_{(i)} - x_{(j)})^2/\phi_j^2 & j > k \end{cases}$$

On the other hand,

$$\begin{aligned}
h_1(\epsilon, \beta) = &\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})\epsilon_i}{\phi_j}\Big)^2 + 2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)} - z)(x_{(i)} - x_{(j)})}{\phi_j}\tilde{\beta}\Big) \\
&- \Big(\frac{\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\epsilon_i}{\phi_k}\Big)^2 + 2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)} - z)(x_{(i)} - x_{(k)})}{\phi_k}\tilde{\beta}\Big)
\end{aligned}$$

First notice that:

$$
\begin{aligned}
&\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)^2-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)^2\\
&=\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)^2-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)^2\\
&=(\phi_k\phi_j)^{-2}\Big(\phi_k^2\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)^2-\phi_j^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2\Big)\\
&=(\phi_k\phi_j)^{-2}\Big(\phi_k^2\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)}+x_{(k)}-x_{(j)})\epsilon_i\Big)^2-\phi_j^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2\Big)\\
&=(\phi_k\phi_j)^{-2}\Big(\phi_k^2\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2+\phi_k^2\Big((x_{(k)}-x_{(j)})\sum_{i=j+1}^{n}\epsilon_i\Big)^2\\
&+2\phi_k^2\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\Big((x_{(k)}-x_{(j)})\sum_{i=j+1}^{n}\epsilon_i\Big)-\phi_j^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2\Big)\\
&=(\phi_k\phi_j)^{-2}\Big(\phi_k^2\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2-\phi_j^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2\Big)+o(1/n)
\end{aligned}
$$

The last term

$$
\begin{aligned}
&\phi_j^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2=\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2\\
&=\left(\phi_k^2+2(x_{(k)}-x_{(j)})\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})+(n-k)(x_{(k)}-x_{(j)})^2+\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})^2\right)\\
&\cdot\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2
\end{aligned}
$$

$$
\begin{aligned}
&\phi_k^2\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)^2-\phi_k^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2\\
&=\phi_k^2\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i-\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i+\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\\
&=\phi_k^2\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i+2\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\\
&=\phi_k^2\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)^2+\phi_k^2\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(2\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)
\end{aligned}
$$

As a result,

$$\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)^2-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)^2$$
$$=(\phi_k\phi_j)^{-2}\left(\phi_k^2\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)^2+2\phi_k^2\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)-\right.$$
$$\left.-\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2\right)$$
$$=o(1)$$

The last equality follows as long as $|k-j|=o(n)$

$$2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)$$
$$-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)$$
$$=2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)$$
$$-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_k}\tilde{\beta}\Big)+O(1/n)$$
$$=2\tilde{\beta}(\phi_j)^{-2}\left(\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)-\phi_j^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\right)$$
$$=2\tilde{\beta}(\phi_j)^{-2}\left(\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)-\phi_j^2\right)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)$$
$$+2\tilde{\beta}(\phi_j)^{-2}\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)$$

$$
\begin{aligned}
&\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)-\phi_j^2\\
&=\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})-\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})^2\\
&=\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})-\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})^2-\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})^2\\
&=\sum_{i=k+1}^{n}(x_{(j)}-x_{(k)})(x_{(i)}-x_{(j)})-\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})^2\\
&=(x_{(j)}-x_{(k)})\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})-\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})^2
\end{aligned}
$$

As a result,

$$
\begin{aligned}
&2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)\\
&-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)\\
&=2\tilde{\beta}(\phi_j)^{-2}\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)\\
&-2\tilde{\beta}(\phi_j)^{-2}\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)+o(1)\\
&=2\tilde{\beta}(\phi_j)^{-2}\Big(\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)+o(1)
\end{aligned}
$$

The last equality follows as long as $|k-j|=o(\sqrt{n})$

Similarly, when $j>k$,

$$
\begin{aligned}
h_2(\epsilon,\beta)=&\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)^2+2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)\\
&-\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)^2+2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)
\end{aligned}
$$

$$
\begin{aligned}
&2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)\\
&-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)\\
&=2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)\\
&-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)+O(1/n)\\
&=2\tilde{\beta}(\phi_j)^{-2}\left(\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)-\phi_j^2\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\right)\\
&=2\tilde{\beta}(\phi_j)^{-2}\left(\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)-\phi_j^2\right)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\\
&-2\tilde{\beta}(\phi_j)^{-2}\Big(\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)
\end{aligned}
$$

$$
\begin{aligned}
&\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)-\phi_j^2\\
&=\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})-\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})^2\\
&=\sum_{i=j+1}^{n}(x_{(j)}-x_{(k)})(x_{(i)}-x_{(j)})\\
&=(x_{(j)}-x_{(k)})\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})
\end{aligned}
$$

As a result,

$$
\begin{aligned}
&2\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i}{\phi_j}\Big)\Big(\frac{\sum_{i=j+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})}{\phi_j}\tilde{\beta}\Big)\\
&-2\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i}{\phi_k}\Big)\Big(\frac{\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(k)})}{\phi_k}\tilde{\beta}\Big)\\
&=-2\tilde{\beta}(\phi_j)^{-2}\Big(\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)+o(1)
\end{aligned}
$$

To summarize,

$$h_1(\epsilon,\beta) = 2\phi_k^2\tilde{\beta}\Big(\sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})\epsilon_i\Big)/\phi_j^2 + o(1)$$

$$h_2(\epsilon,\beta) = -2\phi_k^2\tilde{\beta}\Big(\sum_{i=k+1}^{j}(x_{(i)} - x_{(j)})\epsilon_i\Big)/\phi_j^2 + o(1)$$

$$
\begin{aligned}
P(|\hat{k}-k|>C) &= P(\sup_{|j-k|>C} V(j) \geq V(k)) \\
&\leq P(\sup_{|j-k|>C} |h(\epsilon,\beta)| \geq \inf_{|j-k|>C} \gamma(j)) \\
&\leq P\left(\sup_{|j-k|>C} \left|\frac{h(\epsilon,\beta)}{j-k}\right| \geq \inf_{|j-k|>C} \left|\frac{\gamma(j)}{j-k}\right|\right) \\
&\leq P\left(\sup_{|j-k|>C} \left|\frac{h(\epsilon,\beta)}{j-k}\right| \geq \gamma_K\right)
\end{aligned}
$$

where

$$
\gamma_K = \inf_{|j-k|>C} \begin{cases} \tilde{\beta}^2(\phi_k^2/\phi_j^2)\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})^2/(k-j) & j \leq k \\ \tilde{\beta}^2(\phi_k^2/\phi_j^2)\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})^2/(j-k) & j > k \end{cases}
$$

and since

$$
\begin{aligned}
\frac{h_1(\epsilon,\beta)}{k-j} &= 2\tilde{\beta}(\phi_k^2/\phi_j^2)\Big(\frac{\sum_{i=j+1}^{k}(x_{(i)}-x_{(j)})\epsilon_i}{k-j}\Big) = o(1) \\
\frac{h_2(\epsilon,\beta)}{j-k} &= -2\tilde{\beta}(\phi_k^2/\phi_j^2)\Big(\frac{\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})\epsilon_i}{j-k}\Big) = o(1)
\end{aligned}
$$

Thus, for any ξ, we can find a C such that $P(|\hat{k}-k|>C) < \xi$.

For Asymptotic, consider $\tilde{\beta}_n \to 0$ but $\sqrt{n}\tilde{\beta}_n \to \infty$. In this case, $\gamma(j)$ can no longer be treated as $O_p(1)$ but $O_p(\tilde{\beta}_n^2)$. Thus, the convergence of $\hat{k}$ is

$$
\hat{k} = k + O_p(||\tilde{\beta}_n||^{-2})
$$

Let

$$
\hat{k} = k + v\lambda_n^{-2}
$$

where $\lambda = O_p(||\tilde{\beta}_n||)$ and v is a real number in a compact set. Let

$$K(B) = \{w : w = k + [v\lambda_n^{-2}], |v| < B\}$$

we derive the limiting process of $V(w) - V(k)$ for $w \in K(B)$ and then use continuous mapping theorem for arg max to derive the asymptotic distribution.

$$V(j) - V(k) = \begin{cases} -\tilde{\beta}^2\phi_k^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2/\phi_j^2 + 2\tilde{\beta}(\phi_k^2/\phi_j^2)\Big(\sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})\epsilon_i\Big) & j \le k \\ -\tilde{\beta}^2\phi_k^2 \sum_{i=k+1}^{j}(x_{(i)} - x_{(j)})^2/\phi_j^2 - 2\tilde{\beta}(\phi_k^2/\phi_j^2)\Big(\sum_{i=k+1}^{j}(x_{(i)} - x_{(j)})\epsilon_i\Big) & j > k \end{cases}$$

when $j \le k$

$$\begin{aligned} & -\tilde{\beta}^2 \sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2 + 2\tilde{\beta}\Big(\sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})\epsilon_i\Big) \\ &= -\frac{\tilde{\beta}^2 - v\lambda_n^{-2}}{k-j}\sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2 + 2\frac{\tilde{\beta}\sqrt{-v}\lambda_n^{-1}}{\sqrt{k-j}}\Big(\sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})\epsilon_i\Big) \\ &= vT_1 + 2W_1(-v) \end{aligned}$$

Suppose X is defined on a compact set $[x^-, x^+]$. Then $\sum_{i=j+1}^{k}(x_{(i)} - x_{(j)})^2/(k-j) \to \int_{x^-}^{x_k}(u - x^-)^2 f(u)du$ and define $T_1 = \int_{x^-}^{x_k}(u - x^-)^2 f(u)du$. $W_1(v)$ is a winner process.

when $j > k$

$$
\begin{aligned}
&-\tilde{\beta}^2 \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2 - 2\tilde{\beta}\Big(\sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})\epsilon_i \Big) \\
&= -\frac{\tilde{\beta}^2(v)\lambda_n^{-2}}{j-k} \sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})^2 - 2\frac{\tilde{\beta}\sqrt{v}\lambda_n^{-1}}{\sqrt{j-k}}\Big(\sum_{i=k+1}^{j} (x_{(i)} - x_{(j)})\epsilon_i \Big) \\
&= -vT_2 + 2W_2(v)
\end{aligned}
$$

$\sum_{i=k+1}^{j}(x_{(i)} - x_{(j)})^2/(k-j) \to \int_{x_k}^{x^+}(u-x^+)^2 f(u)du$ and define $T_2 = \int_{x_k}^{x^+}(u-x^+)^2 f(u)du$. $W_2(-v)$ is another winner process.

B.1.8 Proof of Corollary3

Assume $W_\sigma < \infty$ and assume $\sigma(x)$ has bounded second derivatives on X. The density $f(x)$ is a Lipschtiz function. If there exists ω such that for all $m > m_0$, $|\omega_m| > \omega$. Then

$$IMSE_n(m) \le 2\phi_m^2 + 2W_\sigma^2 \frac{K_m}{n} + 8W_1^2 W_\sigma^2 \frac{\log p_n}{n} s_0/\omega^2$$

When the error term involve heteroskedacticity, the choice of λ need to be revised. Define $\Delta = diag(\sigma(x_1), \sigma(x_2), \cdots, \sigma(x_n))$. Define $W_\sigma = \max_x |\sigma(x)|$.

from van de geer book,

$$||M_m D_n(\hat{\psi} - \psi)||_2^2/n + \lambda||\hat{\psi}||_1 \le 2(\Delta\epsilon)' M_m D_n(\hat{\psi} - \psi) + \lambda||\psi||_1$$

$$|(\Delta\epsilon)' M_m D_n(\hat{\psi} - \psi)| \le \Big(\max_{k \le p_n} |(\Delta\epsilon)' M_m D_n^{(k)}| \Big)||\hat{\psi} - \psi||_1$$

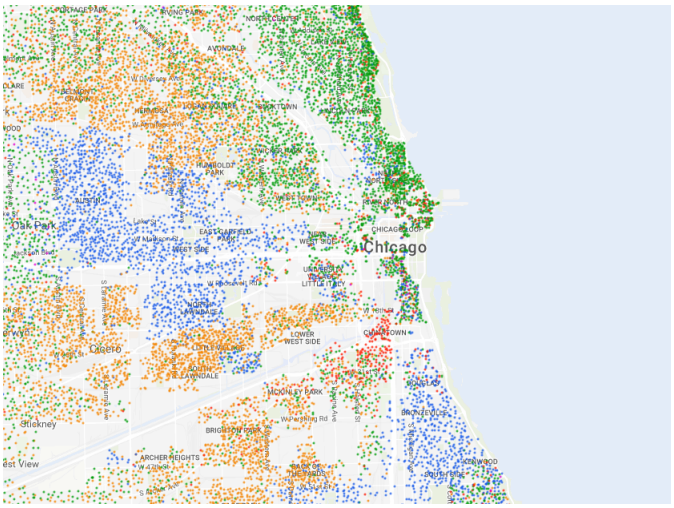
we want to choose λ to overrule the empirical process. Write $X_k = M_m D_n^{(k)}$

$$
\begin{aligned}
P(\max_{k\le p_n} |X_k'\Delta\epsilon|^2 \ge 2\log p_n W_\sigma^2 n) &\le \sum_{i=1}^{p_n} P(|X_k'\Delta\epsilon|^2 \ge 2\log p_n W_\sigma^2 n) \\
&\le \sum_{i=1}^{p_n} 2P(X_k'\Delta\epsilon \ge \sqrt{2\log p_n W_\sigma^2 n}) \\
&\le \sum_{i=1}^{p_n} 2P(Z \ge \frac{\sqrt{2\log p_n W_\sigma^2 n}}{\sqrt{\sum_{l=1}^n x_{kl}^2 \sigma(x_{kl}^2)}}) \\
&\le \sum_{i=1}^{p_n} 2P(Z \ge \frac{\sqrt{2\log p_n W_\sigma^2 n}}{\sqrt{\sum_{l=1}^n x_{kl}^2 W_\sigma^2)}}) \qquad \text{(B.14)} \\
&\le \sum_{i=1}^{p_n} 2P(Z \ge \sqrt{2\log p_n}) \\
&\le \sum_{i=1}^{p_n} \frac{1}{p_n\sqrt{2\log p_n}\sqrt{2\pi}} = \frac{1}{\sqrt{2\log p_n}\sqrt{2\pi}} \\
&\to 0 \quad \text{as} \quad n\to\infty
\end{aligned}
$$

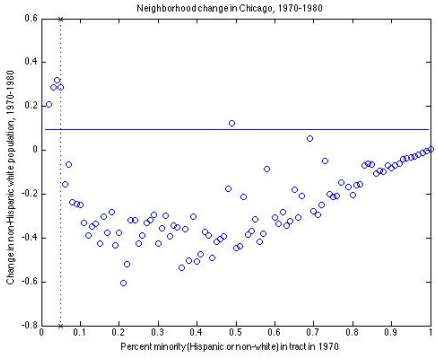
Thus the event $\left\{\max_{k\le p_n} 2|(\Delta\epsilon)' M_n D_n(\hat{\psi}-\psi)| \le \lambda\right\}$ has probability 1 if $\lambda \propto W_\sigma\sqrt{\frac{\log p}{n}}$

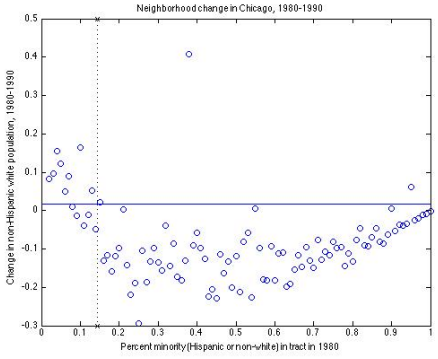
B.2 Figures

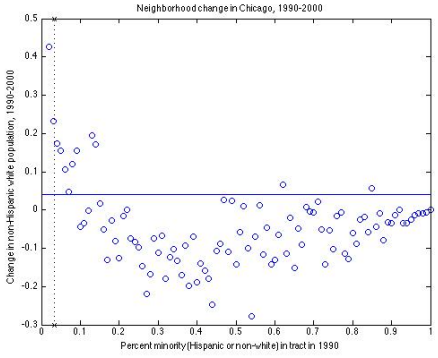
Segregation Map – Chicago

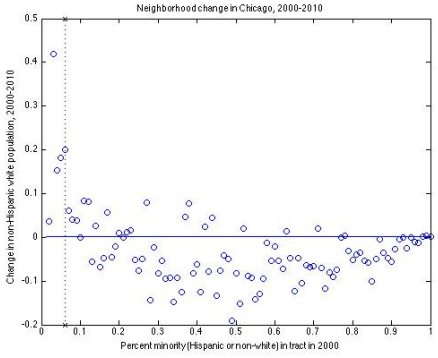


source: [1]

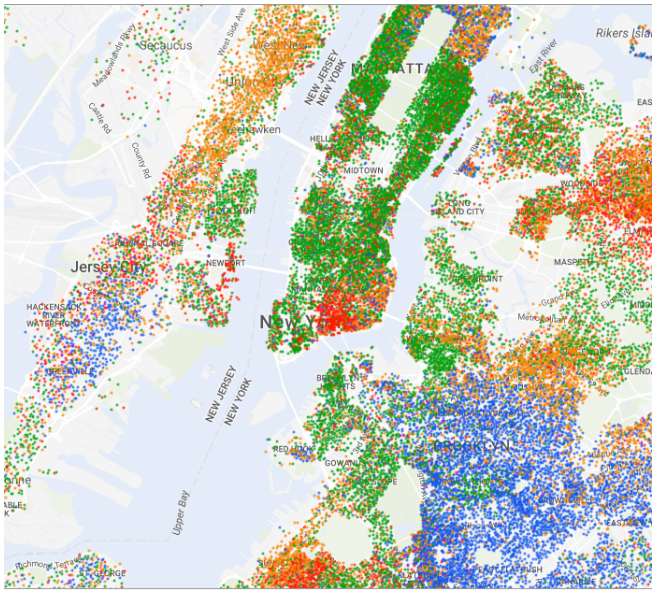




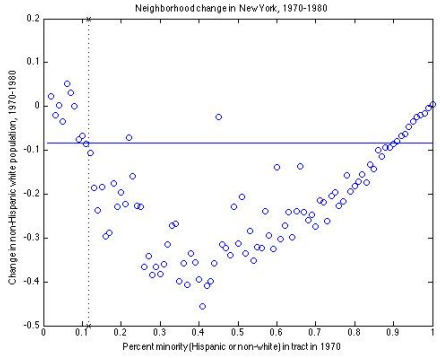


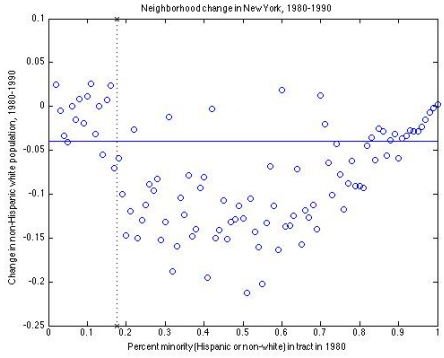


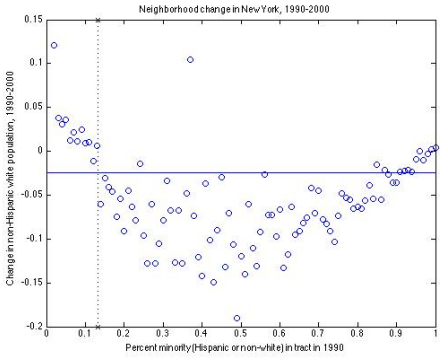
Segregation Map – New York

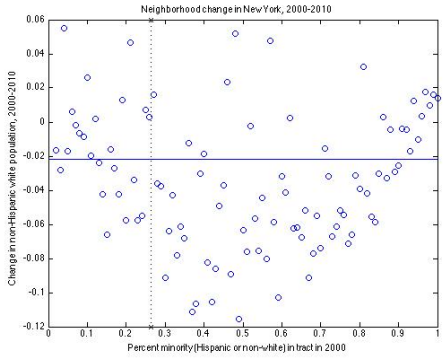


source: [1]

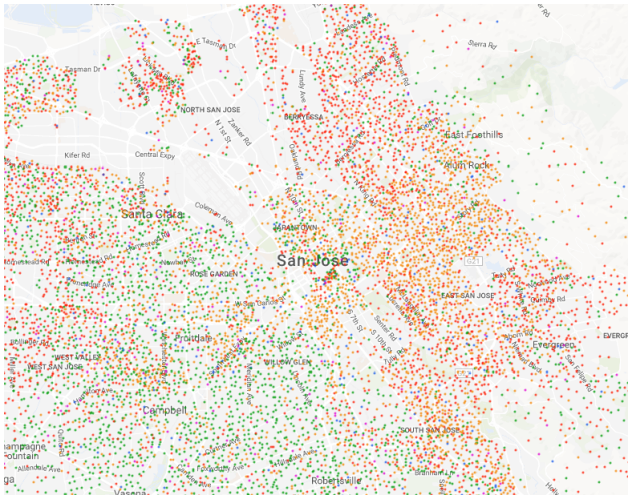




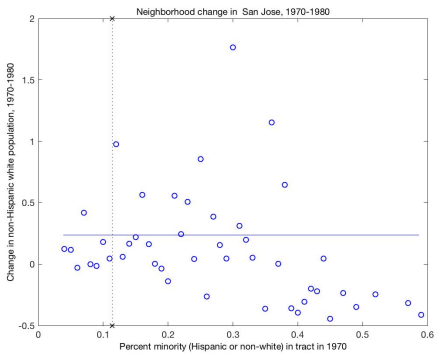


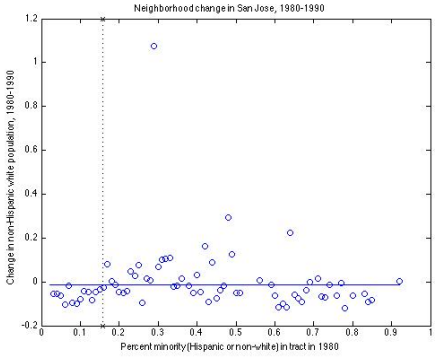


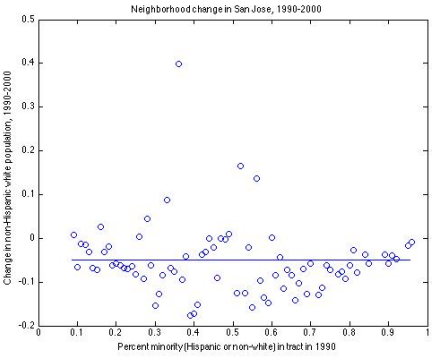
Segregation Map – San Jose

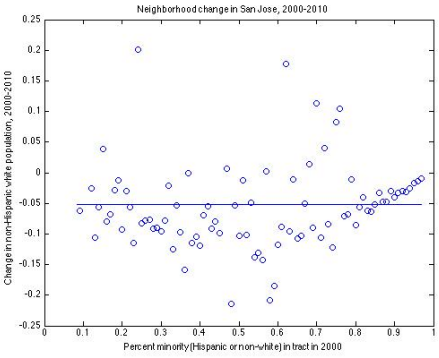


source: [1]









APPENDIX C

CHAPTER 3 OF APPENDIX

C.1 Appendix: Deterministic design points

When the design points $\boldsymbol{x}_i = x_i$, $i = 1, \ldots, n$, are deterministic[1], (3.10) turns into

$$b^2_{x_0}(v) = \left(\sum_{i=1}^{n} \ell_i (s(M(x_i), v) - s(M(x_0), v)) + \sum_{i=1}^{n} \ell_i^- w(M(x_i), v) \right)^2 . \tag{C.1}$$

Since $K(\cdot)$ has compact support in $[-c_K, c_K]$, we have $\ell_i = 0$ if $|x_i - x_0| > c_K h_n$. It is easy to see that all weights are non-negative if and only if

$$\sum K(\frac{x_i - x_0}{h_n}) \left(\frac{x_i - x_0}{h_n} \right)^2 \geq \left| \sum K(\frac{x_i - x_0}{h_n}) \frac{x_i - x_0}{h_n} \right| .$$

This assumption means that the sample rescaled around each point to lie in the range $[-1, 1]$ has the variance that dominates the absolute value of the expectation. For this, the rescaled points should be sufficiently balanced on the left and on the right of x_0. The assumption can be alternatively expressed as

$$\frac{s_2}{h_n^3} \geq c_K \left| \frac{s_1}{h_n^2} \right| .$$

It holds when $s_1/h_n^2 \to 0$ as $n \to \infty$.

By a direct computation, it is possible to show that, in the regular design case, the weights are nonnegative for all n.

Proposition 2. *Consider the local linear setting with uniform kernel supported on* $[-c_K, c_K]$ *and equally spaced (regular) design points* $x_1, \ldots, x_n$ *on a bounded interval* I. *If* $1/n \leq c_K h_n \leq 1$, *then* $\ell_i(x_0) \geq 0$ *for all* i, n *and each*

$$x_0 \in I_n = \{x \in I : \ [x - c_K h_n, x + c_K h_n] \subset I\}.$$

[1]Because with deterministic design $\boldsymbol{x}_i = x_i, i = 1, \ldots, n$, $\boldsymbol{s}_j$, $j = 0, 1, 2$ and $\boldsymbol{\kappa}_{in}, i = 1, \ldots, n$ are also deterministic and we write $\boldsymbol{s}_j = s_j$ and $\boldsymbol{\kappa}_{in} = \kappa_{in}$.

In case of deterministic design points in a bounded interval I, the following assumption is often imposed; they appear as (LP1)-(LP2) in [87].

Assumption M (Design points)**.** *The design points $x_1, \ldots, x_n$ are such that:*

(i) There exists $\lambda_0 > 0$ such that all eigenvalues of $\mathcal{B}_{nx_0}$ are greater than or equal to λ_0 for all sufficiently large n and all $x_0 \in I$.

(ii) There exists $a_0 > 0$ such that, for any interval $J \subset I$ and all $n > 1$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \in J} \leq a_0 \max(\mathrm{Leb}(J)/\mathrm{Leb}(I), 1/n),$$

where $\mathrm{Leb}(\cdot)$ *denotes the Lebesgue measure.*

We impose the following assumption on the response function.

Assumption N (Theoretical response function)**.** *The function $M(x)$, $x \in I$, is defined on a bounded closed interval $I \subset \mathbb{R}$, and there exists $\gamma > 0$ such that, for all $v \in \mathbb{S}^{d-1}$, the derivative of $s(M(x), v)$ with respect to x is Lipschitz with constant γ.*

The following result is similar to [87, Prop. 1.13] in the singleton-valued data framework.

Proposition 3. *If $x_0 \in I_n$, $\ell_i \geq 0$ for all i, and Assumptions I, J, M and N are satisfied, then*

$$|b_{x_0}(v)| \leq c_K^2 C_* \gamma h_n^2, \qquad \sigma_{x_0}^2(v) \leq \frac{\sigma_{\max}^2 C_*^2}{n h_n}$$

for sufficiently large n and $h_n \geq 1/(2n)$.

Proposition 3 implies

$$\mathrm{MSE}(x_0) \leq c_K^4 C_*^2 \gamma^2 h_n^4 + \frac{\sigma_{\max}^2 C_*^2}{n h_n}.$$

Therefore, the upper bound is minimised for the bandwidth given by

$$h_n^* = \left(\frac{\sigma_{\max}^2}{4c_K^4 \gamma^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$

and the following result holds.

Theorem 14. *If the bandwidth is chosen to be $h_n = \alpha n^{-\frac{1}{5}}$ for $\alpha > 0$ and Assumptions I, J, M hold, then*

$$\limsup_{n\to\infty} \sup_{x_0 \in I_n} \mathbb{E}[n^{\frac{2}{5}} L(\hat{M}(x), M(x))] \leq C_1 < \infty,$$

uniformly over all response functions satisfying Assumption N, where C_1 is a constant depending only on γ, a_0, λ_0, $\sigma_{\max}^2$, $K_{\max}$ and α.

C.2 Appendix: Local constant setting

In the local constant case, the weights $\ell_i = \boldsymbol{\kappa}_{in}/(n\boldsymbol{s}_0)$ are always non-negative. Then the estimator $\hat{M}(x_0)$ can be constructed as the convex set whose support functions is obtained by calculating the Nadaraya–Watson estimator for the sample $s(\boldsymbol{Y}_i, v)$, $i = 1, \ldots, n$, in each particular direction v. In other words, $\hat{M}(x_0)$ is the sum of the observed sets $\boldsymbol{Y}_i$ multiplied by non-negative coefficients ℓ_i. Therefore, the bias and variance of the set-valued local constant estimator can be obtained similarly to the singleton-valued data case. For this, it suffices to assume that the function $s(M(x), v)$ is Lipschitz in x with the same constant for all v, which is equivalent to requiring that $M(x)$, $x \in I$, is Lipschitz in the Hausdorff metric.

BIBLIOGRAPHY

[1] Mapping segregation. https://www.nytimes.com/interactive/2015/07/08/us/census-race-map.html?_r=0. Accessed: 2017-04-23.

[2] Daron Acemoglu, Camilo García-Jimeno, and James A. Robinson. Finding eldorado: Slavery and long-run development in colombia. NBER WORKING PAPER SERIES, June 2012.

[3] Andreas Ammermuller and Jörn-Steffen Pischke. Peer effects in european primary schools: Evidence from pirls. *Journal of Labor Economics*, 27(3):315–348, 2009.

[4] Donald W. K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993.

[5] Luc Anselin. *Spatial Econometrics: Methods and Models*. Boston: Kluwer, 1988.

[6] Matt Backus, Thomas Blake, and Steven Tadelis. Cheap talk, round numbers, and the economics of negotiation. NBER Working Paper No. 21285, 2015.

[7] Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. Who's who in networks. wanted: The key player. *Econometrica*, (74):1403–1417, 2006.

[8] Oriana Bandiera, Iwan Barankay, and Imran Rasul. Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica*, 77(4):1047–1094, 2009.

[9] Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, and Matthew Jackson. The diffusion of microfinance. *Science*, 341(6144), 2013.

[10] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.

[11] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[12] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post selection inference for lad regression and other z-estimation problems. Working Paper, 2014.

[13] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, pages 1–18, 2011.

[14] Yoav Benhamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (methodological)*, 57(1):289–300, 1995.

[15] A. Beresteanu and F. Molinari. Asymptotic properties for a class of partially identified models. *Econometrica*, 76:763–814, 2008.

[16] Lawrence E. Blume, William A. Brock, Steven N. Durlauf, and Rajshri Jayaraman. Linear social interactions models. *Journal of Political Economy*, 123(2):444–496, 2015.

[17] Pietro Bonaldi, Ali Hortacsu, and Jakub Kastl. An empirical analysis of funding costs spillovers in the euro-zone with application to systemic risk. 2015.

[18] Ch. Bontemps, T. Magnac, and E. Maurin. Set identified linear models. *Econometrica*, 80:1129–1155, 2012.

[19] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009.

[20] Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 41(2):802–837, 2013.

[21] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[22] Florentina Bunea, Johannes Lederer, and Yiyuan She. The square root group lasso: theoretical properties and fast algorithms. *IEEE-Information Theory*, 60:1313–1325, 2014.

[23] Antoni Calvó-Armengol, Eleonora Patacchini, and Yves Zenou. Peer effects and social networks in education. *Review of Economic Studies*, (76):1239–1267, 2009.

[24] David Card, Alexandre Mas, and Jesse Rothstein. Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123(1):177–218, 2008.

[25] Arun Chandrasekhar, Victor Chernozhukov, Francesca Molinari, and Paul Schrimpf. Inference for best linear approximations to set identified functions. CeMMAP Working Paper CWP 43/12, 2012.

[26] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7:649–688, 2015.

[27] Andrew E. Clark and Youenn Loheac. "it wasn't me, it was them!" social influence in risky behavior by adolescents. *Journal of Health Economics*, 26:763–784, 2007.

[28] A.D. Cliff and John Keith Ord. *Spatial autocorrelation*. London: Pion, 1973.

[29] Tim Coelli, Sanzidur Rahman, and Colin Thirtle. Technical, allocative, cost and scale efficiencies in bangladesh rice cultivation: A nonparametric approach. *Journal of Agricultural Economics*, 53(3):607–626, 2002 2002.

[30] Timothy G. Conley and Christopher R. Udry. Learning about a new technology: Pineapple in ghana. *AMERICAN ECONOMIC REVIEW*, 100(1):35–69, 2010.

[31] Inés Couso and Didier Dubois. Statistical reasoning with set-valued information: ontic vs. epistemic views. *Internat. J. Approx. Reason.*, 55(7):1502–1518, 2014.

[32] Noel A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., 1993.

[33] Aureo de Paula, Imran Rasul, and Pedro CL Souza. Estimating and identifying social interactions. August 2015.

[34] Edward Denbee, Christian Julliard, Ye Li, and Kathy Yuan. Network risk and key players: A structural analysis of interbank liquidity. 2015.

[35] Phil Diamond. Least squares fitting of compact set-valued data. *J. Math. Anal. Appl.*, 147(2):351–362, 1990.

[36] Stephen Donald and Whitney K. Newey. Series estimation of semilinear models. *Journal of Multivariate Analysis*, 50(1):30–40, 1994.

[37] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[38] J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996.

[39] Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, 21(1):196–216, 1993.

[40] Jianqing Fan and Irène Gijbels. Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, 20(4):2008–2036, 1992.

[41] Jianqing Fan and Yuan Liao. Endogeneity in high dimensions. *The Annals of Statistics*, 42(3):872–917, 2014.

[42] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. Working Paper, 2015.

[43] Eric Gautier and Alexandre B. TsyBakov. High-dimensional instrumental variables regression and confidence sets. 2014.

[44] M. A. Gil, M. T. López-García, M. A. Lubiano, and M. Montenegro. Regression and correlation analyses of a linear relation between random intervals. *Test*, 10:183–201, 2001.

[45] Gil González-Rodríguez, Ángela Blanco, Norberto Corral, and Ana Colubi. Least squares estimation of linear regression models for convex compact random sets. *Adv. Data Anal. Classif.*, 1(1):67–81, 2007.

[46] Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. Working Paper, 2015.

[47] Jonathan Guryan, Kory Kroft, and Matthew J Notowidigdo. Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, 1(4):34–68, 2009.

[48] Bruce E. Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000.

[49] Bruce E. Hansen. *Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press, 2014.

[50] Bruce E. Hansen. Regression kink with an unknown threshold. Working Paper, 2015.

[51] D. L. Hawkins. A test for a change point in a parametric model based on a maximal wald-type statistic. *Sankhyā: The Indian Journal of Statistics*, 49(3):368–376, 1987.

[52] William C. Horracea, Xiaodong Liu, and Eleonora Patacchini. Endogenous network production functions with selectivity. *Journal of Econometrics*, 190(2):222–232, 2016.

[53] Fei Jin and Lung-Fei Lee. Lasso maximum likelihood estimation of parametric models with singular information matrices. 2016.

[54] F. Thomas Juster and Richard Suzman. An overview of the health and retirement study. *Journal of Human Resources*, 30 (Supplement):S7–S56, 1995.

[55] Hiroaki Kaido. Asymptotically efficient estimation of weighted average derivatives with an interval censored variable. *Econometric Theory*, 2016. Forthcoming.

[56] Maximilian Kasy. Uniformity and the delta method. *arXiv preprint arXiv:1507.05731*, 2015.

[57] Harry H. Kelejian and Ingmar R. Prucha. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, 17(1):99–121, 1998.

[58] Harry H Kelejian and Ingmar R Prucha. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40:509–533, 1999.

[59] Brian V. Krauth. Peer effects and selection effects on smoking among canadian youth. *Canadian Journal of Economics*, 38(3):735–757, 2005.

[60] David S. Lee. Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142:675–697, 2008.

[61] Lung-fei Lee and Jihai Yu. A spatial dynamic panel data model with both time and individual effects. *Econometric Theory*, 26:564–597, 2010.

[62] Lungfei Lee. Consistency and efficiency of least squares estimation for mixed regressive, spatial. *Econometric Theory*, 18(2):252–277, Apr 2002.

[63] Lungfei Lee. Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive. *Econometric Reviews*, 22(4):305–335, 2003.

[64] Lungfei Lee. Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica*, 72:1899–1926, 2004.

[65] Lungfei Lee and Xiaodong Liu. Efficient gmm estimation of high order spatial autoregrssive models with autoregressive disturbances. *Econometric Theory*, 26:187–230, Feb 2010.

[66] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59, 2005.

[67] Hannes Leeb and Benedikt M. Potscher. Can one estimate the unconditional distribution of post-modelselection estimators? *Econometric Theory*, 24(2):38–376, 2008.

[68] Hannes Leeb and Benedikt M. Potscher. Model selection. *Handbook of Financial Time Series*, pages 889–925, 2009.

[69] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[70] Ye Luo and Victor Chernozhukov. Selecting informative moments via lasso. 2016.

[71] T. Maatouk. *Some application of nonparametric regression with constrained data*. PhD thesis, University of Glasgow, Glasgow, September 2003.

[72] Elena Manresa. Estimating the structure of social interactions using panel data. 2013.

[73] C. F. Manski and E. Tamer. Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70:519–546, 2002.

[74] Charles Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, Jul 1993.

[75] Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Verlag, New York, 2003.

[76] Alexandre Mas and Enrico Moretti. Peers at work. *American Economic Review*, 99(1):112–145, 2009.

[77] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(1436-1462), 2006.

[78] I. Molchanov. *Theory of Random Sets*. Springer, London, 2005.

[79] Ryo Nakajima. Measuring peer effects on youth smoking behaviour. *The Review of Economic Studies*, 74(3):897–935, 2007.

[80] Matthew Neidell and Jane Waldfogel. Cognitive and noncognitive peer effects in early education. *Review of Economics and Statistics*, 92(3):562–576, 2010.

[81] Carmen M. Reinhart and Kenneth S. Rogoff. Growth in a time of debt. *American Economic Review: Papers and Proceedings*, 100:573–578, 2010.

[82] Bruce Sacerdote. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704, 2001.

[83] Georg Schollmeyer and Thomas Augustin. Statistical modeling under partial identification: distinguishing three types of identification regions in regression analysis with interval data. *Internat. J. Approx. Reason.*, 56(part B):224–248, 2015.

[84] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. The sparse group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, May 2013.

[85] Beatriz Sinova, Ana Colubi, María Ángeles Gil, and Gil González-Rodríguez. Interval arithmetic-based simple linear regression between interval data: discussion and sensitivity analysis on the choice of the metric. *Inform. Sci.*, 199:109–124, 2012.

[86] Xiaoying Tian and Jonathan Taylor. Asymptotics of selective inference. August 2016.

[87] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[88] Graham Upton and Bernard Fingleton. *Spatial data analysis by example. Volume 1: Point pattern and quantitative data.* John Wiley and Sons Ltd., 1985.

[89] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

[90] R. A. Vitale. l_p metrics for compact, convex sets. *Journal of Approximation Theory*, 45:280–287, 1985.

[91] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, B(68):49–67, 2006.

[92] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society*, 76(1):217–242, 2011.

[93] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[94] Ying Zhu. Sparse linear models and l1regularized 2sls with high-dimensional endogenous regressors and instruments. 2016.